UNIVERSIDADE FEEVALE

DIEISON MEDINGER

VISUALIZAÇÃO DE DADOS DE PACIENTES COM CÂNCER A PARTIR DA BASE DE DADOS DO INCA.

Novo Hamburgo, Junho de 2017.

DIEISON MEDINGER

VISUALIZAÇÃO DE DADOS DE PACIENTES COM CÂNCER A PARTIR DA BASE DE DADOS DO INCA.

Universidade Feevale
Instituto de Ciências Exatas e Tecnológicas
Curso de Sistemas de Informação
Trabalho de Conclusão de Curso

Orientador: Juliano Varella de Carvalho

Novo Hamburgo, Junho de 2017.

AGRADECIMENTOS

Gostaria de agradecer a todos os que, de alguma maneira, contribuíram para a realização desse trabalho de conclusão, em especial:

Aos meus pais e irmã, pelo incentivo, apoio e amor incondicional. Por me ensinarem grandes virtudes como paciência e perseverança e com muito carinho e apoio, não mediram esforços para que eu chegasse até esta etapa da minha vida.

A minha namorada que apesar de todas as dificuldades me fortaleceu e que para mim foi muito importante.

Aos amigos e companheiros de trabalhos e irmãos na amizade que fizeram parte da minha formação e que vão continuar presentes em minha vida com certeza.

Ao meu orientador Juliano Varella, por todo o suporte, dedicação, suas correções e incentivos no decorrer de todo o trabalho.

RESUMO

Câncer é uma das doenças que mais intriga cientistas e pesquisadores por não haver uma cura 100% eficaz e nem causas comprovadas, podendo variar de caso para caso. Estimativas para 2016 e 2017 apontam 596.070 casos novos de câncer, sendo 49% em mulheres e 51% em homens, reforçando a magnitude do problema no país. Estudos e pesquisas ligadas a esta doença vêm ganhando mais espaço com o passar dos anos, não só na esfera nacional, mas também mundial. Em todo o mundo existem entidades e ONGs (Organizações não governamentais) que buscam fomentar discussões, estudos e promover fóruns sobre o tema. Assim, na Europa há o *Cancer Research UK*, nos Estados Unidos da America o *American Cancer Society* que incentivam e promovem pesquisas sobre a doença. No Brasil há o INCA (Instituto Nacional de Câncer) que além de investimentos para área de pesquisa disponibiliza uma base de dados sobre pacientes, com informações de idade, sexo, estado, município de nascimento e residência, bem como onde o paciente foi diagnosticado com câncer, tipo de tumor, histórico de alcoolismo e tabagismo, entre outros. Esta base mantém registros entre os anos de 1985 e 2015. Em um primeiro momento, para profissionais da área de tecnologia da informação, esta base de dados pode ser de grande valia, demonstrando um potencial enorme para exploração dos dados ali contidos através da construção de gráficos e visualizações customizadas, busca de padrões através de técnicas de *data mining* ou até mesmo um ponto de partida para comparação dos casos ali contidos com bases de institutos de outros países. Porém, quando esta base de dados, de forma crua, é apresentada a um profissional da área da saúde, ele não tem a mesma capacidade e conhecimento técnico para explorar todo o potencial destes dados como um profissional de TI. Pensando nesta dificuldade que um pesquisador não ligado a área da tecnologia venha a ter para interpretar e utilizar de forma eficiente estas informações disponibilizadas é que propõe-se o desenvolvimento deste trabalho. A fim de gerar visualizações e gráficos customizados, foi criada uma ferramenta para visualização dessa base de uma forma onde o usuário pode interagir com os dados e investigar a informação ali contida, também aplicando regressão linear, uma técnica de mineração de dados, agregando assim novas perspectivas aos dados já existentes. Foram geradas visualizações interativas, gráficos atraentes e instigantes com a capacidade de também explorar regressões lineares. Utilizou-se recursos e bibliotecas disponíveis na linguagem de programação R. Buscando facilitar a disponibilização destes dados, todo o conteúdo gerado foi disponibilizado de forma online, em um servidor R rodando uma aplicação Shiny. Após a finalização da ferramenta alguns profissionais da área foram convidados a realizar testes e deixar *feedback* quanto a facilidade de uso e relevância desta nova solução para a comunidade de pesquisadores.

Palavras-chave: Gráficos. *Data Mining*. Câncer. Visualização de dados. Descoberta de Conhecimento.

ABSTRACT

Cancer is a disease that most intrigues scientists and researchers because there is no 100% effective cure and no proven causes, which may vary from case to case. Estimates for 2016 and 2017 show 596,070 new cases of cancer, 49% in women and 51% in men, reinforcing the magnitude of the problem in the country. Studies and researches related to this disease are gaining more space over the years, not only at national level but also worldwide. All around the world there are entities and NGOs (non-governmental organizations) that seek to foster discussion, promote studies and forums on the subject. As in Europe has Cancer Research UK, in the United States of America has American Cancer Society that encourage and promote research on the disease. In Brazil, there is the INCA (Instituto Nacional de Câncer) that in addition to investments in research area has a database of patients with information on age, sex, state, county of birth and residence, and where the patient has been diagnosed with cancer, tumor type, history of alcoholism and smoking, among others. This base keeps records between 1985 and 2015. At first, for professionals in the information technology area, this database can be of great value, demonstrating an enormous potential for exploitation patterns through data mining techniques or even a starting point for comparison of cases there contained with institute's bases from other countries. However, when this database, crudely, is presented to a professional in the health field, he does not have the same capacity and technical knowledge to exploit the full potential of these data as an IT professional. Thinking about this difficulty that a researcher not connected to the area of technology will have to interpret and make efficient use of this information is that it proposes the development of this work. In order to generate visualization and custom graphs, it has been created a tool to visualize this database where the user can interact with the data and investigate the information, also generating linear regression, a datamining technique, adding new perspectives to the existing data. Interactive visualizations, attractive and thought-provoking graphics with the ability to also explore linear regressions have been generated. It has been used available resources and libraries in the programming language R. In order to facilitate the availability of this data, all generated content was made available online, on an R server running a Shiny application. After completing the tool, some professionals in the field were invited to perform tests and leave feedback on the ease of use and relevance of this new solution to the research community.

Key words: Graphs. Data Mining. Cancer. Data Visualization. Knowledge Discovery.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

WHO	World Health Organization
INCA	Instituto Nacional de Câncer
IARC	Internal Agency for Research on Cancer
PIB	Produto Interno Bruto
IDC	International Data Corporation
KDD	Knowledge Discovery in Database
GSP	Generalized Sequential Pattern
DHP	Direct Hashing and Pruning
DIC	Dynamic Itemset Counting
SAS	Statical Analysis System

SUMÁRIO

INTRODUÇÃO

Câncer é um termo genérico para um vasto grupo de doenças que podem afetar qualquer parte do corpo. Assim como câncer há outros termos para descrever esta doença como tumores malignos ou neoplasias malignas. Uma característica marcante desta doença é o crescimento descontrolado de células a partir de mutações genéticas. O câncer pode ocorrer em quase qualquer órgão ou tecido do corpo, assim como pode se disseminar para além dos limites do órgão ou tecido de origem, processo conhecido como metástase, maior causador das mortes por câncer. (INCA, 2016)

Além de ser uma doença extremamente agressiva, os números de casos e mortes é um dos fatores que mais assusta a comunidade de cientistas e pesquisadores desta área da saúde. Conforme mostra o portal da World Health Organization (2016), 8,2 milhões de pessoas morrem todos os anos pela doença, o que representa 13% de todas as mortes no mundo. É esperado um aumento de 70% nos casos da doença nas próximas 2 décadas, sendo que atualmente já foram diagnosticados mais de 100 tipos diferentes de câncer, e cada tipo requer diagnóstico e tratamento único.(WHO,2016)

Além dos dados mundiais alarmantes, no Brasil estas estatísticas não são diferentes. Conforme estimativas divulgadas pelo Instituto Nacional de Câncer (INCA), para 2016-2017 são esperados cerca de 420 mil casos novos de câncer dos mais variados tipos e em torno de mais 180 mil incidências de câncer de pele.

Levando em conta os dados apresentados acima, é possível ver que este é um problema de âmbito global. Muitas entidades apoiam e financiam pesquisas, estudos e debates sobre o tema. Entidades internacionais como *World Health Organization (WHO), American Cancer Society*, *UK Cancer Society* e outras nacionais como o Instituto Oncoguia, Fundação do Câncer e Instituto Nacional de Câncer fornecem dados, estatísticas e até mesmo *datasets* com as mais variadas informações sobre pacientes com históricos da doença.

Como visto, informações relevantes aos pesquisadores que estudam esta doença são encontradas de várias maneiras nas mais variadas fontes e

disponibilizadas em diferentes formatos. Tan, Steinbach e Kumar (2006), afirmam que muitas vezes não é possível usar técnicas tradicionais para analisar dados, mesmo em pequenos conjuntos de dados, necessitando assim desenvolver novos métodos para efetuar esta análise.

Porém, devido à natureza da formação de pesquisadores da área da saúde, compilar, manipular e analisar, de forma eficiente estes dados pode ser uma difícil tarefa. Pensando nesta dificuldade o atual trabalho propõe o uso da informática e suas técnicas para analisar informações coletadas sobre a doença, além de gerar visualizações que estimulem a criatividade e o senso de investigação dos mesmos.

Coleman e Ahlemeyer-Stubbe (2014) escrevem que a habilidade de extrair conhecimento oculto em dados tem se tornado muito importante no mundo atual. Quando os dados são usados para predição e comportamentos futuros isso representa uma grande vantagem na área em que estes dados são estudados.

Buscando-se uma maneira de extrair essas informações dos dados, de maneira mais acurada e utilizando técnicas eficazes, optou-se por usar a linguagem R. Além de ser uma linguagem de uso livre e código aberto, esta apresenta bibliotecas que atendem a todas as necessidades deste trabalho, como por exemplo a biblioteca Shiny, que provê solução para construir aplicações web interativas. A ligação automática "reativa" entre entradas e saídas e extensos *widgets* pré-construídos tornam possível a construção de belos aplicativos, responsivos e poderosos com o mínimo de esforço. (CHANG *et al*., 2017)

O pacote shiny ainda prevê uma aplicação que pode ser instalada em um servidor Linux e disponibiliza as aplicações para usuários externos acessarem como se fosse um site externo. Esta solução permite gerenciar processos R sendo executados em URLs e portas diferentes com outras vantagens como hospedar várias aplicações simultaneamente.

O trabalho gerou visualizações, dinâmicas e estáticas, capazes de instigar a investigação dos pesquisadores, dando-lhes a liberdade de manipular

os gráficos gerados. Para isso foram investigadas várias bibliotecas da linguagem R.

A pesquisa de métodos e bibliotecas para representar os dados graficamente se faz de extrema importância para o resultado final deste trabalho. Conforme Chen, Härdle e Unwin (2008) a visualização de dados é um termo novo, e expressa mais do que a representação dos dados em forma de gráfico, ela deve ajudar os leitores a ver a estrutura dos dados, assim como deduzir informações sobre estes dados em vez de apenas concentrar-se na apresentação de informações.

> Ferramentas de visualização ajudam as pessoas em situações em que vendo a estrutura do *dataset* em detalhe é melhor do que ver apenas um breve resumo do mesmo. Uma dessas situações ocorre quando explorar os dados para encontrar padrões, tanto para confirmar os padrões esperados quanto encontrar aqueles inesperados. Outra situação ocorre quando se avalia a validade de um modelo estatístico, para julgar se o modelo de fato se ajusta aos dados. (MUNZNER, 2014, p. 7, Tradução do Autor)

"A evolução do poder computacional tem sido de grande benefício para a geração de gráficos nos últimos anos. Tornou-se possível desenhar precisos e complexos gráficos com grande facilidade e imprimi-los com uma qualidade impressionante em alta resolução." (CHEN, HÄRDLE E UNWIN, 2008, p.5, tradução nossa). Mesmo encontrando-se gráficos com as mais variadas informações sobre o câncer na internet, poucos exploram de forma eficiente o poder da computação ou possibilitam a interação e manipulação das visualizações fornecidas ao público que acessa o conteúdo.

De forma mais pobre e rudimentar são encontradas visualizações sobre esta doença com dados e estatísticas no âmbito nacional no portal do INCA. Este portal fornece algumas visualizações onde se é possível ter alguma interação com os gráficos ali apresentados.

Ao final do desenvolvimento deste trabalho, buscou-se obter uma ferramenta online que, além de consolidar informação sobre incidências de câncer em uma única plataforma, ainda utilizou de recursos gráficos para exibir informações sobre a doença de maneira interativa e com possibilidade de manipular estes gráficos em tempo de execução.

Este trabalho foi dividido em quatro capítulos, o primeiro trata do assunto câncer, apresentando-se estatísticas e dados sobre a doença. O segundo aborda conceitos de mineração de dados e técnicas que podem ser utilizadas. O terceiro apresenta técnicas utilizadas para visualizar dados, assim como informações históricas do surgimento e necessidade de se visualizar dados através de gráficos. O quarto e último capítulo trata da construção da ferramenta, dificuldades e soluções encontradas para conclusão da aplicação. Finalmente, as considerações finais e as referências bibliográficas.

1. CÂNCER

O câncer é o nome dado a uma coleção de doenças que estão intimamente relacionadas. Em todos os tipos de câncer existe uma característica comum, a divisão desordenada de algumas células do corpo, e o possível espalhamento para órgãos e tecidos, além daquele onde a doença foi originada. (INCA,2016)

O câncer pode começar quase que em qualquer parte do corpo humano, o qual é constituído por trilhões de células. Normalmente, as células humanas crescem e se dividem para formar novas células conforme o corpo precisa delas. Quando as células envelhecem ou ficam danificadas, elas morrem e novas células tomam o seu lugar. Esse processo é chamado de apoptose, ou morte celular programada, que ajuda na homeostasia do corpo humano. Quando o câncer se desenvolve, no entanto, este processo ordenado é quebrado. (INCA,2016)

Essa quebra no processo fisiológico faz com que células anormais sobrevivam. Assim, o processo neoplásico inicia em uma única célula, chamada de célula-mãe, que é capaz de transmitir às células filhas as mesmas características alteradas. A alteração ocorre no DNA por mutação e essa alteração genética é transmitida para as células filhas por mitose. Estas células podem dividir-se desordenadamente, além de crescer além dos parâmetros de normalidade, resultando na formação de tumores. (INCA,2016)

1.1 A doença

O câncer não é uma doença moderna. Desde os primórdios da história muitos autores e estudiosos têm escrito sobre a doença. Como descrito pela American Cancer Society (2014) algumas das primeiras evidências de câncer foram encontradas em tumores ósseos fossilizados no antigo Egito. Sendo que o registro mais antigo encontrado foi em 3000 a.C. conhecido como Edwin Smith Papiro, e se trata da cópia de um antigo livro egípcio que descreve 8 casos de tumores ou úlceras das mamas que foram removidos por cauterização com uma ferramenta chamada broca de fogo. Porém neste manuscrito não foi utilizada a palavra câncer, mas foi descrito que não há nenhum tratamento para a doença.

> O termo “câncer” (do latim câncer = caranguejo) é a tradução latina da palavra grega *Karkinos* (crustáceo, caranguejo). Foi usado pela primeira vez por GALENO (aproximadamente 138-201 d.C) para designar um tumor maligno da mama cuja veias superficiais apareciam inturgescidas e ramificadas, lembrando as patas de um caranguejo. O emprego da palavra então generalizou-se para indicar tumores malignos de qualquer natureza. (TEIXEIRA *et al*., 1997, p.13)

Autores divergem quanto ao surgimento do termo câncer e dos primeiros registros da doença no antigo Egito. Como descreve Sudhakar (2009, tradução nossa) “A palavra câncer provinha da palavra grega Karkinos que era utilizada para descrever tumores de carcinoma empregada por um médico chamado Hipócrates (460-370 a.C.) ”, este mesmo autor também relata o manuscrito dos antigos egípcios como sendo de 1600 a.C.

> O médico romano, Celsus (25 a.C - 50 d.C.), mais tarde traduziu o termo grego em câncer, a palavra latina para caranguejo. Galen (130-200 d.C.), outro médico grego, usou a palavra *oncos* (termo grego para inchaço) para descrever tumores. Embora a analogia de Hipócrates e Celsus ainda é usado para descrever tumores malignos, o termo de Galen é agora utilizada como uma parte do nome para especialistas em câncer - oncologistas. (AMERICAN CANCER SOCIETY, 2014, tradução nossa)

Apesar das divergências entre os autores sobre o uso do termo câncer e da data de suas primeiras incidências, ela é uma doença que atualmente é muito pesquisada, devido a agressividade e altos índices de mortalidade. Uma característica marcante desta doença é o crescimento descontrolado de células que além de atingirem qualquer parte do corpo, multiplicam-se além dos seus limites usuais. Em todos os tipos de câncer as células continuam a crescer e dividir em vez de morrer, formando assim novas células anormais, que podem atingir partes adjacentes do corpo e se espalharem para outros órgãos através da circulação sanguínea ou vasos linfáticos. Esse processo é conhecido como metástase e é o maior causador de mortes de câncer (WHO, 2016).

De acordo com a Encyclopædia Britannica (2016) um tumor pode ser classificado em dois grandes grupos. Quando mantém-se na mesma área em que se originou e apresenta poucos riscos à saúde este é classificado como benigno. Já quando o tumor cresce e se espalha de forma agressiva, este é classificado como maligno.

Porém essa classificação é muito branda, como mostra Encyclopædia Britannica (2016, tradução nossa):

> “Malignos e benignos são distinções importantes, mas eles são categorias gerais que compõem muitas formas diferentes de câncer. Uma forma mais detalhada e útil para classificar e citar os diversos

> tipos de tumores é por seu local de origem (a célula ou tecido a partir do qual surge um tumor) e pela sua aparência microscópica. Esse esquema de classificação, embora não seguido com lógica rígida ou consistência, permite que os tumores sejam categorizados por um comportamento clínico típico, tal como prognóstico, e pela resposta a terapia."

Atualmente, seguindo a classificação de acordo com local de origem, existem mais de 100 tipos de câncer e cada um exige único diagnóstico e tratamento. Quando falamos em causas, no passado haviam muitas teorias sobre o que poderia causar um tumor. Hipócrates acreditava na teoria humoral, ou seja, o corpo continha 4 humores (4 fluidos corporais), sangue, fleuma, bile amarela e bile negra. Quando estes 4 humores estavam balanceados a pessoa estava saudável, porém os desequilíbrios destes líquidos causavam doenças. O excesso de bile negra causava os tumores (SUDHAKAR, 2009).

Muito mais tarde, em torno dos anos de 1800 surgiram algumas famosas teorias na Alemanha. A primeira delas foi proposta por Johannes Muller (1801-1858), chamada de teoria blastema, a qual demonstrou que o câncer é composto de células e não linfa, ele também acreditava que os tumores não vinham de células normais. Muller propôs que as células cancerígenas desenvolviam-se a partir de elementos de brotamento (blastema) entre tecidos normais. Mais tarde, seu aluno Rudolph Virchow (1821-1902), um famoso patologista alemão, sugeriu a teoria da irritação crônica como causadora de câncer, mas incorretamente ele acreditava que o câncer se espalhava como liquido. O que mais tarde foi refutado por Karl Thiersch (1822-1895), que mostrou que a metástase acontece através da disseminação de células malignas e não através de algum liquido conforme havia mencionado Virshow. (AMERICAN CANCER SOCIETY, 2014)

Após esta época, uma nova teoria foi sustentada até os anos de 1920, conhecida como teoria do trauma. Ela resistiu bastante tempo, até novas descobertas e teorias que levaram as causas que conhecemos hoje sobre prováveis causadores de câncer.

Desde 1915, quando a doença foi induzida em coelhos através da utilização de alcatrão de carvão, muito já foi descoberto na pesquisa sobre possíveis causas. (AMERICAN CANCER SOCIETY, 2014) Hoje já foram identificados alguns fatores de risco que podem levar a anomalias no crescimento e multiplicação das células tais como tabaco, álcool, hormônios,

infecções, hereditariedade, radiação solar, exposição a substâncias químicas, entre outros. (NATIONA CANCER INSTITUTE, 2016)

> Um fator de risco é algo que aumenta a chance de uma pessoa de desenvolver câncer. Embora fatores de risco influenciem muitas vezes o desenvolvimento de câncer, a maioria não diretamente causa câncer. Algumas pessoas com vários fatores de risco nunca desenvolvem câncer, enquanto outros sem fatores de risco conhecidos desenvolvem. (CANCER.NET, 2016, tradução nossa)

O *National Cancer Institute* (2016) mostra que não há como saber se uma pessoa vai desenvolver câncer, mas pesquisas recentes apontam que os fatores mencionados acima podem aumentar as chances de se desenvolver tal anomalia. Este instituto ainda aponta que cientistas pesquisaram grandes grupos de pessoas e compararam os que desenvolveram a doença com os que não a desenvolveram. Estes estudos podem mostrar que as pessoas que desenvolvem câncer são mais ou menos propensos a se comportar de determinadas maneiras ao ser expostos a determinadas substâncias do que aqueles que não desenvolvem câncer. (NATIONA CANCER INSTITUTE, 2016)

Da mesma maneira que não há cientificamente comprovado as causas que levam uma pessoa a desenvolver câncer e sim fatores que aumentam o risco, não há uma cura comprovada. “Cirurgiões antigos sabiam que o câncer normalmente volta depois de ser removido por cirurgia. ” (SUDHAKAR, 2009, p.3, tradução nossa) Da mesma maneira que hoje ainda acontece, o que existe são maneiras e tratamentos que possibilitam estender o tempo de vida dos pacientes e melhorar sua qualidade de vida.

Os mais comuns tratamentos do câncer atualmente são cirurgia, quimioterapia e terapia por radiação. Cada técnica pode ser usada sozinha ou em combinação com outras técnicas. A escolha do tratamento depende do tipo de câncer, da extensão da doença, a sua taxa de progressão, do estado do doente, e a resposta à terapia. (SUDHAKAR, 2009)

1.2 Estatísticas e base de dados da doença

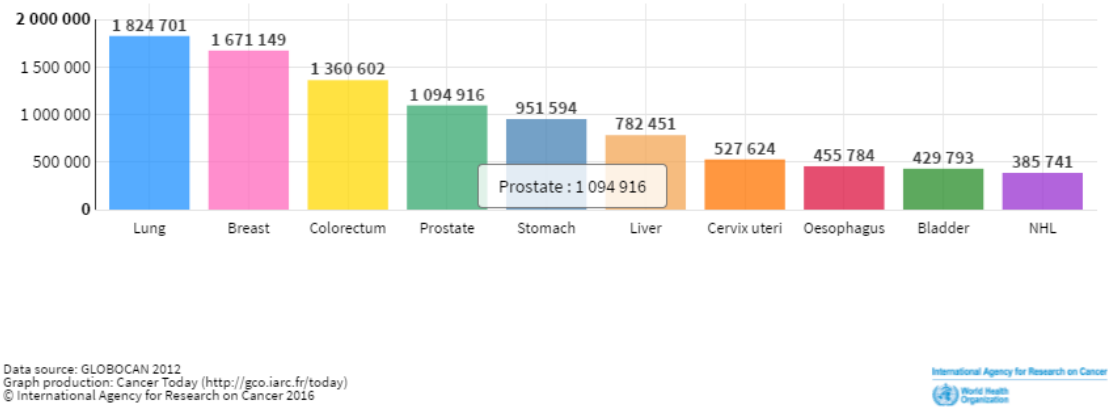
Apesar de ser uma doença extremamente agressiva e possuir apenas fatores de risco que levam ao desenvolvimento de câncer e que podem variar entre indivíduos, outro ponto a ser levado em conta são as estatísticas, números de incidentes e mortes decorrentes desta anomalia.

Com o advento e evolução da internet o acesso a informação acontece de maneira muito fácil, através dos mais variados meios dentro da grande rede. Informações de todos os tipos são geradas e disponibilizadas todos os dias. Dados, informações e estatísticas sobre a doença também são geradas e disponibilizadas por várias entidades, tanto no âmbito global como nacional.

O portal *World Health Organization* (2016) disponibiliza informações sobre a doença. Neste são mostrados números como de 8.2 milhões de pessoas que morrem todos os anos de câncer. Este número representa 13 % do total de mortes no mundo. Nas próximas duas décadas é esperado um aumento de 70% no número de incidentes no âmbito global.

De maneira mais interativa, o portal do *Internal Agency for Research on Cancer* (IARC) contém dados globais da doença para o ano de 2012 apresentados em gráficos como barras, treemap, mapas, entre outros. Conforme mostra a figura 1, gerado neste portal, é possível visualizar o número de casos de câncer separados por tipos. A figura 1 ainda mostra que os dois cânceres mais comum em 2012 foram o câncer de pulmão (Lung) e mama (Breast).

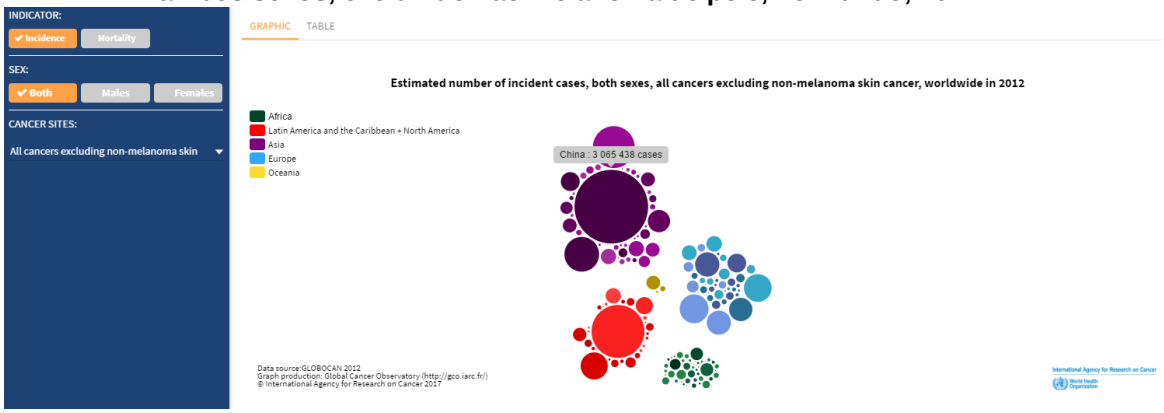
Figura 1 - Número estimado de casos, ambos sexos, no mundo (top 10 tipos de câncer) em 2012



Fonte: http://gco.iarc.fr/today/online-analysis-multi-bars

É possível filtrar algumas informações como sexo, região, incidência ou mortalidade e alguns outros parâmetros conforme o tipo de gráfico selecionado. Conforme mostra o gráfico contido na figura 2, podemos ver que é possível selecionar os dados a serem exibidos por tipo de câncer.

Figura 2 - Gráfico Circle packing com número estimado de incidências de câncer para ambos sexos, excluindo não melanoma de pele, no mundo, 2012.



Fonte: http://gco.iarc.fr/today/online-analysis-circle-packing

A figura 2 mostra a divisão de casos de câncer por região. Cada cor denomina uma região, quando posicionado o mouse sobre um dos círculos pode-se ver o nome do país correspondente. O tamanho do círculo define a quantidade de casos relatados naquele país, ou seja, quanto maior o círculo mais casos de câncer.

Conforme relata Kasper et al. (2015) nos Estados Unidos da América o censo de 2014 apontou que 1.665 milhões de pessoas foram diagnosticadas com câncer, 855.220 homens e 810.320 mulheres. Já 585.720 pessoas morreram da doença no mesmo ano. Câncer é a causa de uma em quatro mortes nos Estados Unidos. A figura 3 mostra a tabela com a divisão entre os tipos de câncer relatados por Kasper.

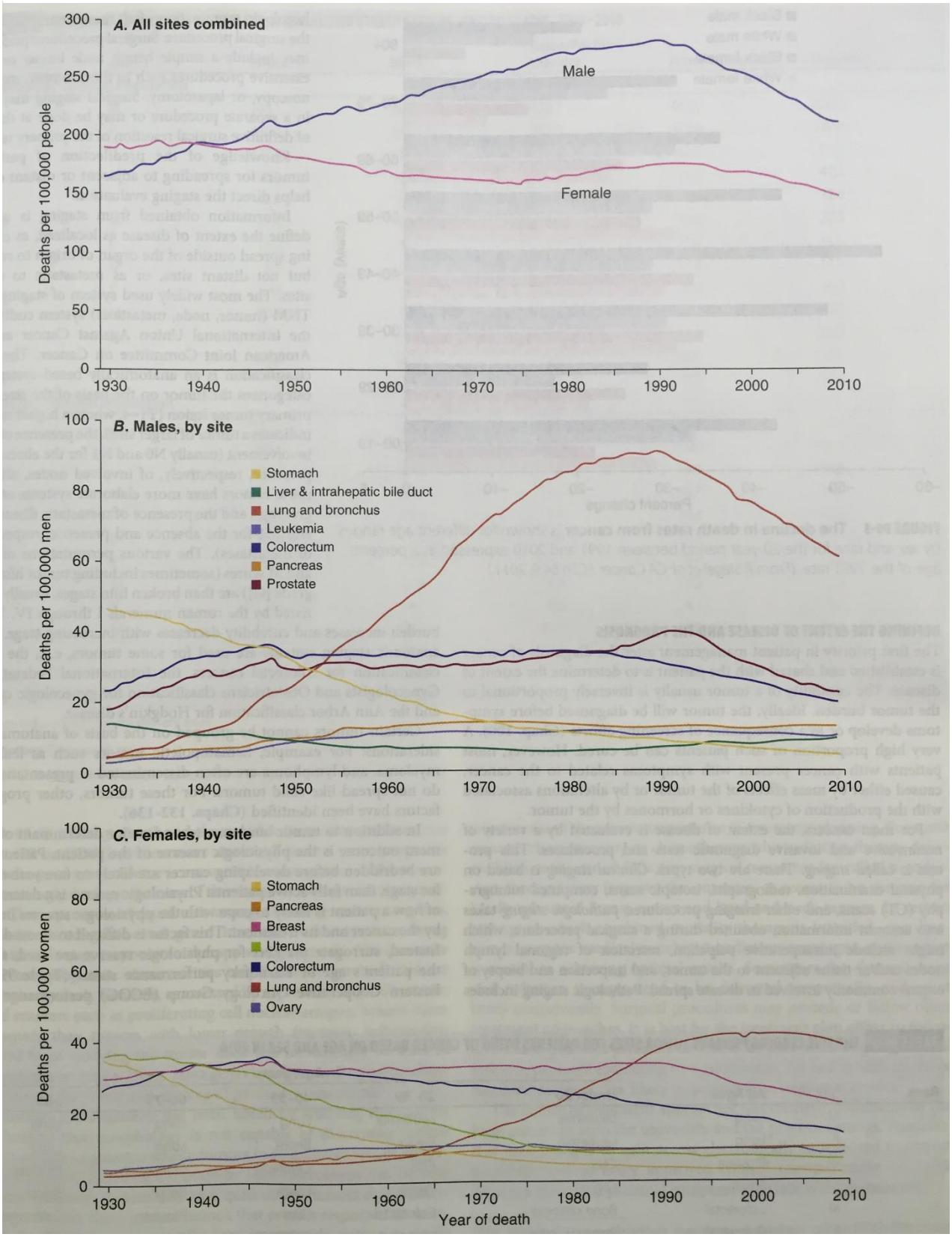
Figura 3 - Distribuição de incidências e mortes por câncer, 2014

	Male			Female	
Sites	**%**	**Number**	**Sites**	**%**	**Number**
Cancer Incidence					
Prostate	27	233,000	Breast	29	232,670
Lung	14	116,000	Lung	13	108,210
Colorectal	8	71,830	Colorectal	8	65,000
Bladder	7	56,390	Endometrial	6	52,630
Melanoma	5	43,890	Thyroid	6	47,790
Kidney	4	39,140	Lymphoma	4	32,530
Lymphoma	4	38,270	Melanoma	4	32,210
Oral cavity	4	30,220	Kidney	3	24,780
Leukemia	4	30,100	Pancreas	3	22,890
Liver	3	24,600	Leukemia	3	22,280
All others	20	171,780	All others	21	169,330
All sites	100	855,220	All sites	100	810,320
Cancer Deaths					
Lung	28	86,930	Lung	26	72,330
Prostate	10	29,480	Breast	15	40,000
Colorectal	8	26,270	Colorectal	9	24,040
Pancreas	7	20,170	Pancreas	7	19,420
Liver	5	15,870	Ovary	5	14,270
Leukemia	5	14,040	Leukemia	4	10,050
Esophagus	4	12,450	Endometrial	3	8,590
Bladder	4	11,170	Lymphoma	3	8,520
Lymphoma	3	10,470	Liver	3	7,130
Kidney	3	8,900	CNS	2	6,230
All others	23	74,260	All others	23	65,130
All sites	100	310,010	All sites	100	275,710

Fonte: KASPER, 2015 p. 467

Além da tabela mostrada na figura 3 Kasper ainda disponibilizam dados sobre o histórico de mortes da doença entre os anos de 1930 e 2010, separados por sexo e pelos tipos de câncer mais comuns em gráficos de linhas que são exibidos na figura 4.

Figura 4 - Gráfico de linhas com dados de mortes por câncer entre 1930 – 2010, EUA.



Fonte: KASPER, 2015 p. 469

Como pode-se observar na Figura 4, alguns tipos de câncer tiveram um

aumento expressivo nas taxas de morte, enquanto outros tiveram uma retração no número total de óbitos. Já no somatório de todos os tipos entre homens e mulheres, observa-se uma curva ascendente que começa em 1930 para homens com pico máximo em 1990 e uma leve regressão até 2010. Já para mulheres a variação no número de mortes por câncer nunca ultrapassou os 210 casos por 100 mil habitantes e apenas após 2005 teve o número de mortes inferior a 150 casos por 100 mil habitantes.

Tabela 1 - Classificação dos 15 primeiros países com maior número de incidências de câncer - 2012

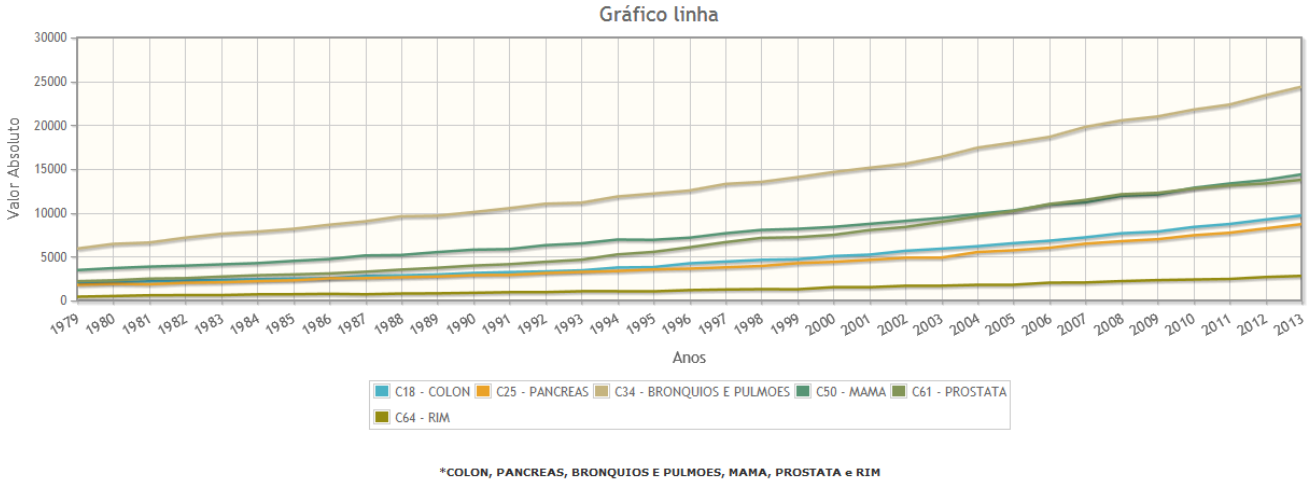
N º	País	Incidência
1	China	3 065 438
2	United States of America	1 603 586
3	India	1 014 934
4	Japan	703 863
5	Germany	493 780
6	Russian Federation	458 382
7	Brazil	437 592
8	Italy	354 456
9	France (metropolitan)	349 426
10	United Kingdom	327 812
11	Indonesia	299 673
12	Korea, Republic of	219 520
13	Spain	215 534
14	Canada	182 182
15	Poland	152 216

Fonte: http://gco.iarc.fr/today

A tabela 1 foi extraída do portal IARC (2016). Foram extraídas apenas as 15 primeiras linhas. Claramente pode-se notar que o câncer não se trata de uma doença que assombra apenas países subdesenvolvidos, é possível ver que as maiores economias do mundo se encontram nas primeiras posições. Porém se a tabela fosse classificada de acordo com a taxa da população que desenvolveu a doença, esta estaria em uma ordem diferente. Estes dados são importantes para planejar futuras ações de prevenção ao câncer. Comparar o número de incidentes por renda per capita, percentagem total da população que contraiu a doença e até mesmo fazer uma relação com o Produto Interno Bruto (PIB) de cada país, pode revelar países ou regiões que precisam de ajuda externa para combater a doença ou até mesmo necessitam de campanhas mais fortes de conscientização para mudança de hábitos.

No Brasil estas estatísticas não são diferentes, conforme estimativas divulgadas pelo INCA para 2016-2017 é esperado cerca de 420 mil casos novos de câncer dos mais variados tipos e em torno de mais 180 mil incidências de câncer de pele. A estimativa aponta 300.870 novos casos em homens e em torno de 295.200 casos para mulheres.

Figura 5 - Total de mortes, por anos, segundo localização primária do tumor, em homens e mulheres, Brasil, entre 1979 e 2013



Fonte: https://mortalidade.inca.gov.br/

A figura 5 foi gerada no portal do INCA (2016) para número de óbitos por câncer. Foi selecionado os tumores de cólon, pâncreas, brônquios e pulmões, mama, próstata e rins entre os anos de 1979 e 2013. O que chama atenção neste gráfico, é que quando comparado com gráfico anterior, da figura 4, com dados referentes a tumores de mama, próstata e pulmões que mais mataram nos últimos anos nos EUA, vemos que as linhas seguem uma alta continua nos números de óbitos diferente da oscilação entre altas e baixas que ocorrem nas últimas décadas nos casos de mortes na América do Norte.

Figura 6 - Estimativas de novos casos de câncer, em mulheres, para 2016/2017

Localização Primária	Casos Novos	%
Mama feminina	57.960	28,1%
Cólon e Reto	17.620	8,6%
Colo do útero	16.340	7,9%
Traqueia, Brônquio e Pulmão	10.890	5,3%
Estômago	7.600	3,7%
Corpo do útero	6.950	3,4%
Ovário	6.150	3,0%
Glândula Tireoide	5.870	2,9%
Linfoma não Hodgkin	5.030	2,4%
Sistema Nervoso Central	4.830	2,3%
Leucemias	4.530	2,2%
Cavidade Oral	4.350	2,1%
Esôfago	2.860	1,4%
Pele Melanoma	2.670	1,3%
Bexiga	2.470	1,2%
Linfoma de Hodgkin	1.010	0,5%
Laringe	990	0,5%
Todas as Neoplasias sem pele*	**205.960**	
Todas as Neoplasias	300.870	

Fonte: http://www.oncoguia.org.br/conteudo/estimativas-no-brasil/1705/1/

Figura 7 - Estimativas de novos casos de câncer, em homens, para 2016/2017

Localização Primária	Casos Novos	%
Próstata	61.200	28,6%
Traqueia, Brônquio e Pulmão	17.330	8,1%
Cólon e Reto	16.660	7,8%
Estômago	12.920	6,0%
Cavidade Oral	11.140	5,2%
Esôfago	7.950	3,7%
Bexiga	7.200	3,4%
Laringe	6.360	3,0%
Leucemias	5.540	2,6%
Sistema Nervoso Central	5.440	2,5%
Linfoma não Hodgkin	5.210	2,4%
Pele Melanoma	3.000	1,4%
Linfoma de Hodgkin	1.460	0,7%
Glândula Tireoide	1.090	0,5%
Todas as Neoplasias sem pele*	**214.350**	
Todas as Neoplasias	295.200	

Fonte: http://www.oncoguia.org.br/conteudo/estimativas-no-brasil/1705/1/

O portal Oncoguia apresentou em 2015 os dados apresentados nas Figuras 6 e 7, com a estimativa para os anos de 2016/2017 contendo o número de novos casos de câncer em homens e mulheres no Brasil. Números que mostram o quanto esta doença merece atenção no país.

Quando analisado tumores específicos para o Brasil, uma análise útil é verificar a taxa de mortalidade atrelada a cada faixa etária. Como exemplo buscou-se a taxa de mortes ligadas a tumores bucais como lábio, base da língua, outras partes da língua, gengiva, assoalho da boca, palato, outras partes da boca, glândula parótida, outras glândulas salivares maiores e amigdala, para então através do portal do INCA (2016) gerar uma tabela separada por faixa etária contendo o número total de óbitos para os anos de 1979 a 2013 como pode ser visto na Figura 8.

Figura 8 - Taxas de mortalidade pelas topografias selecionadas por idade, por 100.000 homens e mulheres, Brasil, entre 1979 e 2013.

	Homens		Mulheres	
Faixa Etária	Número de Obito	Taxa Específica	Número de Obito	Taxa Específica
00 a 04	42	0,01	30	0,01
05 a 09	31	0,01	14	0
10 a 14	27	0,01	17	0,01
15 a 19	46	0,02	43	0,02
20 a 29	261	0,05	191	0,04
30 a 39	1.723	0,44	503	0,12
40 a 49	9.778	3,36	1.515	0,5
50 a 59	17.922	9,15	2.837	1,34
60 a 69	15.236	12,4	3.752	2,69
70 a 79	9.134	14,72	4.217	5,52
80 ou mais	4.332	19,77	4.218	12,83
Idade ignorada	152	8,95	27	1,56
Total	**58.684**	-	**17.364**	-
Taxa Bruta	-	**2,14**	-	**0,62**
Tx Padr. Mundial	-	**2,8**	-	**0,69**
Tx Padr. Brasil	-	**2,93**	-	**0,76**

Fonte: https://mortalidade.inca.gov.br/

Para análises mais acuradas e uso de técnicas computacionais mais avançadas, o INCA ainda disponibiliza de forma aberta uma base de dados sobre pacientes, com informações de idade, sexo, estado, município de nascimento e residência, bem como onde o paciente foi diagnosticado com câncer, tipo de tumor, histórico de alcoolismo e tabagismo, entre outros. Esta base mantém registros entre os anos de 1985 e 2015. Este recurso possibilita, por meio de

técnicas computacionais como mineração de dados, a busca por padrões, comparação de taxa de mortes e incidentes entre anos, tumores, entre outros. Através desta base é possível a geração de gráficos como os exibidos acima com possibilidade de uma melhor manipulação dos dados.

1.3 Considerações finais

O estudo do câncer, atualmente é uma área que vem ganhando mais espaço. Há muitas entidades que apoiam pesquisas ligadas à área. Porem a disponibilização destes dados e principalmente visualizações ainda acontece de maneira muito precária. Conforme dados e gráficos coletados e compilados neste capitulo, podemos ver que para pesquisadores independentes é muito difícil fazer analises com as ferramentas hoje disponíveis.

No âmbito mundial, o portal do IARC fornece uma opção interessante de visualizar dados a respeito da doença, oferecendo algumas opções de filtragem e manipulação de dados, porém, até o momento da confecção deste trabalho, o portal apenas havia disponibilizados dados para o ano de 2012.

Na esfera nacional, o INCA fornece mais tipos de visualizações com maios possibilidades de filtrar estas informações, com dados referentes aos anos de 1985 a 2015 e ainda uma base de dados que pode ser exportada e manipulada nos mais diversos softwares para análise de dados.

Ainda assim é possível notar que há uma carência muito grande quanto a disponibilização de uma ferramenta que possibilite a analise destes dados, consolidando-os em uma única plataforma e fornecendo mais tipos de visualizações que possibilitem a investigação e comparação dos dados.

2. MINERAÇÃO DE DADOS

Os constantes avanços em várias áreas tecnológicas têm impulsionado o aumento na geração, coleta e armazenamento de dados. O aumento de capacidade de armazenamento e processamento de informações, com custo reduzidos também contribuíram para o rápido aumento da quantidade de dados gerada e armazenada.

Porém, possuir um grande volume de dados não significa possuir um grande volume de informações úteis a um dado propósito. Além do grande volume de dados, outro fator que pode tornar essa análise inviável com as técnicas existentes é a natureza dos dados. Para ambas as situações, e nas demais, onde os métodos existentes não respondem as questões necessárias, é imprescindível a criação de novos métodos. Nesse contexto surge a mineração de dados, a fim de suprir essa carência na hora de analisar grandes montantes de dados.

Conforme descrito por Tan, Steinbach e Kumar (2006), a mineração de dados combina métodos tradicionais de análise com algoritmos sofisticados, tornando possível o processamento de grandes volumes de dados de maneira automática. A mineração ainda possibilita explorar e analisar novos tipos de dados e analisar tipos antigos de novas maneiras.

2.1 Descoberta de Conhecimento – KDD

A *International Data Corporation* (IDC) estimou que em 2012 foram gerados 2.8 zettabytes de dados (1 zettabyte é igual a 1 bilhão de terabytes). Mesmo com todo esse montante de dados, antes da era do *Big Data*[1]*,* empresas não davam muito valor aos dados coletados que não gerassem um valor imediato a empresa. Porém, após o início dessa nova era, essa situação se inverteu e as empresas têm investido cada vez mais em coletar e manter informações por seu potencial valor futuro (DEAN, 2014).

Devido a esse aumento no valor da informação, armazená-la e analisá-la da maneira mais eficiente se tornaram tarefas importantes para qualquer

[1] O autor, Jean Dean considera como sendo 2001 o início da era do *Big Data.*

empresa. Não só no mundo corporativo a análise de um grande volume de dados representa geração de valor. Conforme ainda descrito pelo mesmo autor, Dean (2014), o uso de mineração de dados ajudou a identificar que o uso de tamoxifeno não era 80% efetivo nos pacientes, mas sim 100% efetivo em 80% dos pacientes.

> "O tamoxifeno é muitas vezes uma das primeiras drogas prescritas para o tratamento de câncer de mama porque tem uma elevada taxa de sucesso de cerca de 80%. Aprender que uma droga é 80% eficaz nos dá esperança de que tamoxifeno vai proporcionar bons resultados para os pacientes, mas há um importante detalhe sobre a droga que não era conhecido até a era do *Big Data*. É que o tamoxifeno não é 80% eficaz em pacientes, mas 100% eficaz em 80% dos pacientes e ineficaz no resto. Essa é uma constatação de mudança de vida para milhares de pessoas a cada ano." (DEAN, 2014, p.26, tradução nossa)

Na crescente necessidade de extrair conhecimento destas bases de dados é que surgiu o *Knowledge Discovery in Databases* (KDD), que como aponta Fayyad; Piatetsky-Shapiro; Smyth (1996), "em um nível abstrato, o KDD está preocupado com o desenvolvimento de métodos e técnicas para construir sentido aos dados. "

Katoua (2013) ainda destaca a confusão de muitos autores quanto aos termos KDD e *Data Mining*, onde muitos afirmam serem expressões sinônimas, porém a mineração é um elemento da descoberta de conhecimento. "Na primeira conferência internacional de KDD ocorrida em Montreal, 1995, propôs-se que o termo KDD seria empregado para descrever todo o processo de extração de conhecimento a partir de dados. " (KATOUA, 2013, p.2, tradução nossa)

O KDD é caracterizado como um processo composto por várias etapas. Passos e Goldschmidt (2005) descrevem estes processos como pré-processamento compreendendo as funções relacionadas à captação, organização e o tratamento dos dados, os quais têm o objetivo de preparar os dados para os algoritmos da etapa seguinte, a Mineração de Dados. Nesta etapa é realizada a busca efetiva por conhecimentos úteis. A etapa seguinte, de pós-processamento, trata da análise do conhecimento obtido na mineração.

De maneira mais detalhada, Fayyad; Piatetsky-Shapiro; Smyth (1996) descrevem 9 passos no processo de descoberta de conhecimento em banco de dados:

O primeiro consiste em desenvolver uma compreensão do domínio da aplicação e do conhecimento prévio relevante, identificando o objetivo do processo de KDD do ponto de vista do cliente.

Em segundo lugar, selecionar um conjunto, subconjunto de variáveis ou amostras de dados onde a descoberta será executada.

O terceiro passo envolve a limpeza dos dados e pré-processamento. Operações básicas incluem a remoção de ruídos, se necessário, coleta de informações para um modelo, decisão de estratégias para lidar com a ausência de campos nos dados.

Em um quarto passo faz-se a redução dos dados e projeção, encontrar recursos úteis para representar os dados. Com redução de dimensionalidade ou métodos de transformação, o número efetivo de variáveis em consideração pode ser reduzido, ou representações invariáveis para os dados podem ser encontradas.

No quinto é verificado o alinhamento dos objetivos da mineração, traçados no passo um, com um método (ou técnica) como compactação, classificação, regressão, agrupamento, e assim por diante.

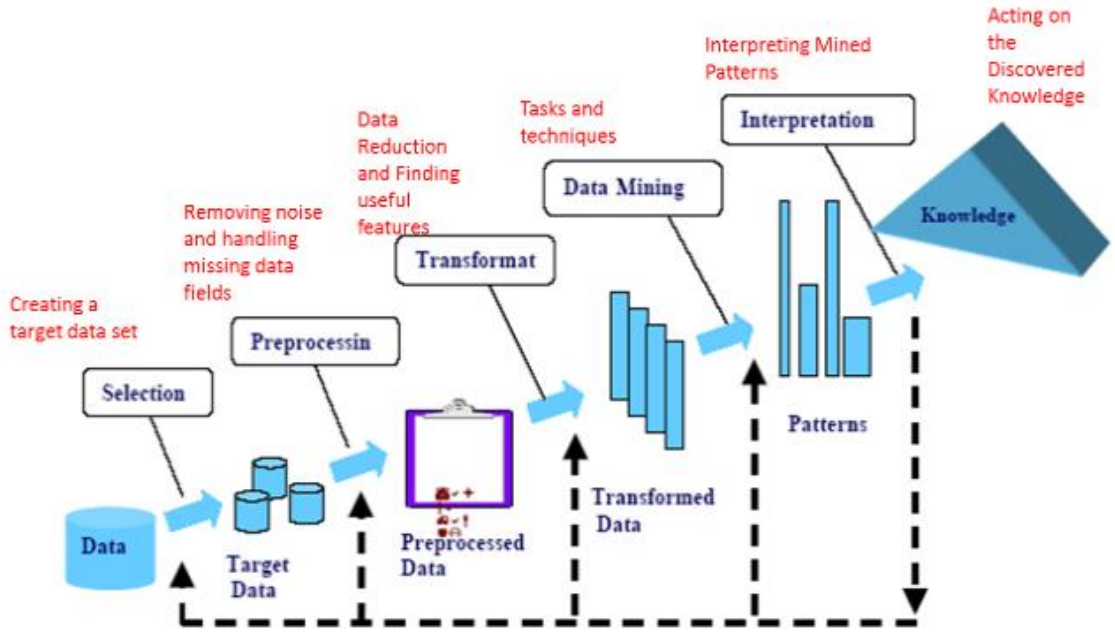
O passo seis é a análise exploratória, modelos e seleção de hipóteses. Neste passo são selecionados os algoritmos e métodos para análise e busca de padrões.

Em um sétimo passo é que a mineração dos dados acontece, a busca por padrões de interesse em uma forma de representação particular ou um conjunto de tais representações, incluindo as regras de classificação ou árvores, regressão e *clustering*, dependendo do método utilizado.

O oitavo passo consiste na interpretação da mineração e, em muitos casos, retornando a qualquer etapa de 1 a 7 para outra iteração. Esta etapa também compreende a visualização dos padrões extraídos e modelos.

No último passo é onde o conhecimento gerado e descoberto é utilizado, incorporando-o em outro sistema ou documentando as descobertas.

Figura 9 - Fases do processo de descoberta de conhecimento.



Fonte: Katoua, 2013 p.44

Conforme mostra a Figura 9, Katoua (2013) considera as 5 etapas descritas por Fayyad: seleção, processamento, transformação, mineração e interpretação, como fundamentais para a descoberta de conhecimento, onde a mineração de dados é o elemento chave para se alcançar este objetivo.

2.2 Técnicas de Mineração de Dados

A mineração de dados é usada para uma variedade de propósitos em ambos os setores: público e privado. Bancos, seguradoras, indústrias médicas e varejistas costumam usar mineração de dados para reduzir custos, estudar a satisfação dos clientes com os produtos e aumentar as vendas.

Devido a esta enorme gama de possibilidades de uso da mineração de dados, é necessário que seja selecionada a técnica adequada ao tipo de resultado esperado e também que se encaixe ao tipo e natureza dos dados que serão minerados.

Atualmente existem muitas técnicas já largamente utilizadas. Katoua (2013) construiu a tabela 2 para demonstrar as principais tarefas e seus algoritmos:

Tabela 2 - Data Mining Tarefas e Técnicas

Tarefa Data Mining	Técnica Apropriada de Data Mining
Classificação	Redes Neurais
	Máquinas de Vetores de Suporte
	Árvores de Decisão
	Algoritmo Genético
	Regra de Indução
Clustering	K-means
Regressão e Predição	Máquinas de Vetores de Suporte
	Árvores de Decisão
	Regra de Indução, NN
Associação e análise de links (Descobrir correlações entre itens em um conjunto de dados)	Regra de mineração por associação
Summarization	Visualização Multivariada

Fonte: Adaptado pelo autor

Há outras opções de tarefas de mineração, porém serão introduzidas neste trabalho aquelas com mais aderência à natureza dos dados que serão utilizados para alcançar os objetivos finais: classificação, associação, regressão e análise de grupos.

2.2.1 Classificação

A tarefa de classificação trata de organizar objetos em uma categoria pré-definida. Tan, Steinbach e Kumar (2006) definem classificação como a tarefa de aprender uma função alvo que mapeie cada conjunto de atributos x para um dos rótulos de classe y pré-determinados. Essa função alvo também é conhecida como modelo de classificação.

A classificação, de acordo com Passos e Goldschmidt (2005), é uma das mais importantes e populares tarefas de mineração de dados. Uma vez identificada essa função, ela pode ser aplicada a novos registros de forma a prever a classe dos mesmos.

Para se aplicar a classificação a um conjunto de dados atualmente existem alguns algoritmos disponíveis como o algoritmo de Hunt, rede Bayesiana de classificação, árvore de decisão, redes neurais, entre outros.

2.2.1.1 Exemplo de classificação

Para exemplificar o funcionamento da tarefa de classificação, usou-se a linguagem R e o *dataset* iris, que faz parte da biblioteca de exemplos contidas no pacote de instalação do R. Algumas outras bibliotecas foram usadas para melhorar a visualização da árvore de decisão construída.

Esse *dataset* é composto por 150 observações contendo 5 variáveis: comprimento e largura da sépala, comprimento e largura das pétalas e espécies de íris (uma espécie de flor).

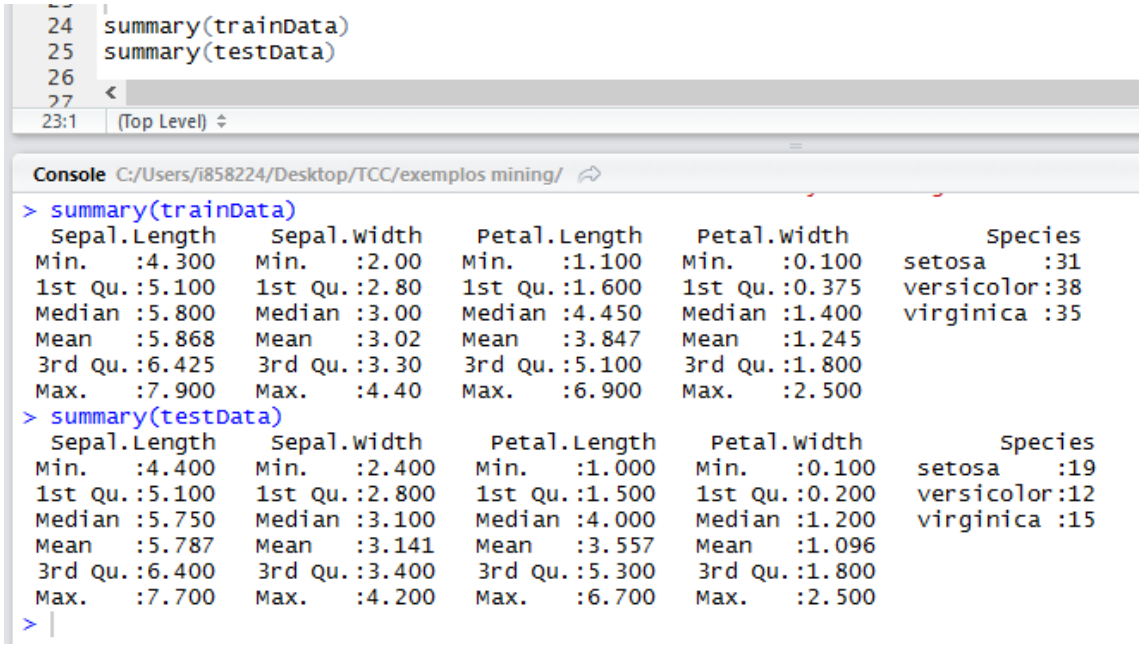
Classificação é uma tarefa supervisionada, onde precisamos de dados pré-classificados e novos dados que serão classificados de acordo com os padrões descobertos nos dados já pré-classificados.

Geralmente, possui-se uma percentagem dos dados disponíveis para testes e outra percentagem para treinamento e a construção de um modelo de classificação. Por exemplo, é preciso primeiro treinar um modelo para detectar quais e-mails são *spam* usando uma base de dados já pré-classificada para então com este modelo verificar futuros *spams*.

Para exemplificar a tarefa de decisão usou-se um algoritmo para construção de uma árvore de decisão. Uma árvore de decisão é uma estrutura em forma de fluxograma, onde cada nó interno denota um teste em um atributo, cada ramo representa um resultado do teste e cada nó de folha (ou nó terminal) contém um rótulo de classe. O nó mais alto de uma árvore é chamado de raiz.

Antes de aplicar o algoritmo *rpart* para criar uma árvore de decisão é necessário separar a massa de dados em duas bases menores. Para o exemplo foi utilizado 70% das observações para compor a base de treinamento e 30% para testar o modelo criado utilizando a base de treinamento. Para esta separação utilizou-se a função *sample* contida na linguagem R, que dividiu a base nas percentagens descritas de forma aleatória.

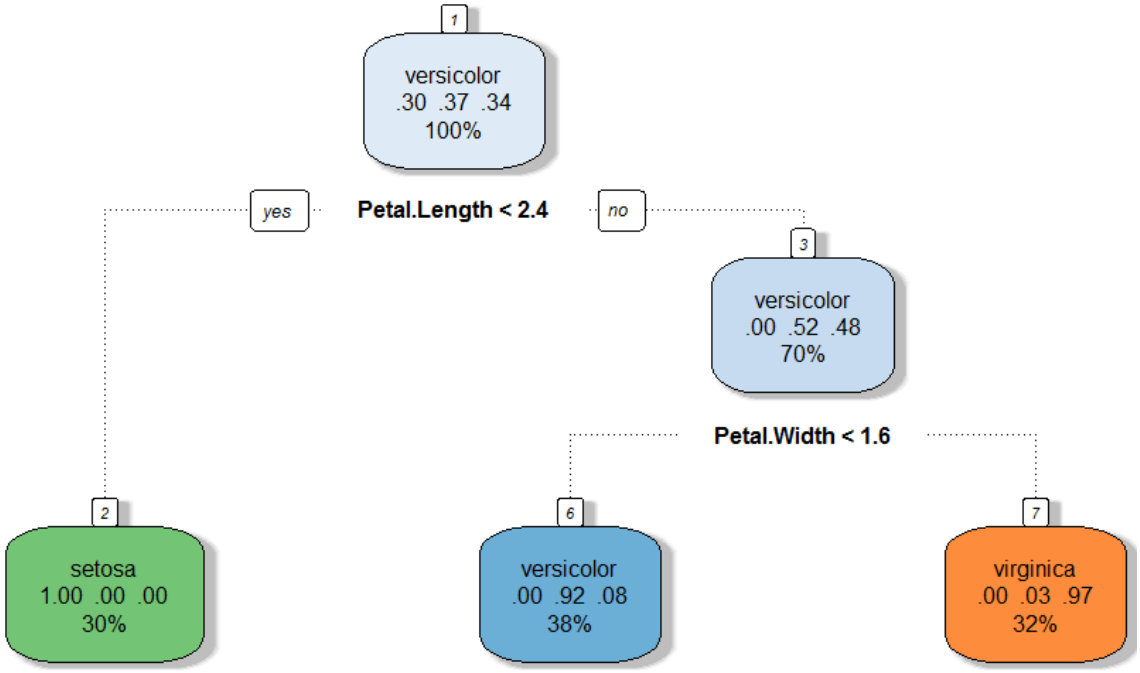
Figura 10 - Composição das bases de treino e teste



Fonte: Adaptado pelo autor

Com a base de treino já definida, utilizou-se o algoritmo *rpart(Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width, data = trainData, method = "class"),* onde busca-se através de análise dos atributos de comprimento e largura da sépala, comprimento e largura das pétalas para cada espécie a fim de construir um modelo que classifique a espécie da base de treino utilizando apenas os dados das sépalas e pétalas das observações das flores íris.

Figura 11 - Árvore de decisão



Fonte: Adaptado pelo autor

A Figura 11 ilustra a árvore de decisão para classificação do tipo de flor íris. O nó raiz, no topo da árvore, mostra a divisão das classes, 30%, 37% e 34% entre *setosa, versicolor* e *virginica* respectivamente e a percentagem das observações que passaram pelo teste, 100% neste primeiro nó. Ou seja, todas as observações foram submetidas a pergunta “Pétala é maior que 2,4 cm?” Se sim essa observação se trata de uma flor íris do tipo Seteosa, se não a observação é submetida ao próximo teste no nó seguinte. As observações passam por todos os testes até serem classificadas. Esse processo se repete até que todas as observações sejam classificadas. Nos nós seguintes da figura 11 vemos os testes executados e a classificação de cada grupo.

Após a construção do modelo de predição, utilizou-se essa árvore de decisão para classificar a base de teste. A Figura 12 mostra o resultado após submeter a base de teste ao modelo de árvore de decisão construído. A primeira coluna contém as espécies já classificadas na base de dados original, a segunda a classificação encontrada pela árvore de decisão. Pode-se ver que a árvore de decisão criada acertou em torno de 95%, um número considerado satisfatório.

Como visto no exemplo de classificação, essa tarefa pode gerar um método para classificar elementos de forma muito eficiente. No exemplo usou-se a árvore de decisão com o algoritmo *rpart.* Porém dependendo o número de

variáveis avaliadas e a natureza dos dados analisados outros algoritmos podem ter um desempenho melhor.

Figura 12 - Resultado da classificação de espécies de Íris

```
> testData[,5:7]
        Species SpecieClass SpecieProb.setosa SpecieProb.versicolor SpecieProb.virginica
1        setosa      setosa        1.00000000             0.00000000            0.00000000
6        setosa      setosa        1.00000000             0.00000000            0.00000000
7        setosa      setosa        1.00000000             0.00000000            0.00000000
10       setosa      setosa        1.00000000             0.00000000            0.00000000
12       setosa      setosa        1.00000000             0.00000000            0.00000000
13       setosa      setosa        1.00000000             0.00000000            0.00000000
17       setosa      setosa        1.00000000             0.00000000            0.00000000
19       setosa      setosa        1.00000000             0.00000000            0.00000000
23       setosa      setosa        1.00000000             0.00000000            0.00000000
24       setosa      setosa        1.00000000             0.00000000            0.00000000
33       setosa      setosa        1.00000000             0.00000000            0.00000000
34       setosa      setosa        1.00000000             0.00000000            0.00000000
35       setosa      setosa        1.00000000             0.00000000            0.00000000
36       setosa      setosa        1.00000000             0.00000000            0.00000000
39       setosa      setosa        1.00000000             0.00000000            0.00000000
40       setosa      setosa        1.00000000             0.00000000            0.00000000
45       setosa      setosa        1.00000000             0.00000000            0.00000000
47       setosa      setosa        1.00000000             0.00000000            0.00000000
50       setosa      setosa        1.00000000             0.00000000            0.00000000
56   versicolor  versicolor        0.00000000             0.92500000            0.07500000
64   versicolor  versicolor        0.00000000             0.92500000            0.07500000
68   versicolor  versicolor        0.00000000             0.92500000            0.07500000
75   versicolor  versicolor        0.00000000             0.92500000            0.07500000
78   versicolor   virginica        0.00000000             0.03030303            0.96969697
80   versicolor  versicolor        0.00000000             0.92500000            0.07500000
82   versicolor  versicolor        0.00000000             0.92500000            0.07500000
83   versicolor  versicolor        0.00000000             0.92500000            0.07500000
86   versicolor  versicolor        0.00000000             0.92500000            0.07500000
90   versicolor  versicolor        0.00000000             0.92500000            0.07500000
93   versicolor  versicolor        0.00000000             0.92500000            0.07500000
98   versicolor  versicolor        0.00000000             0.92500000            0.07500000
101   virginica   virginica        0.00000000             0.03030303            0.96969697
103   virginica   virginica        0.00000000             0.03030303            0.96969697
105   virginica   virginica        0.00000000             0.03030303            0.96969697
109   virginica   virginica        0.00000000             0.03030303            0.96969697
112   virginica   virginica        0.00000000             0.03030303            0.96969697
116   virginica   virginica        0.00000000             0.03030303            0.96969697
121   virginica   virginica        0.00000000             0.03030303            0.96969697
123   virginica   virginica        0.00000000             0.03030303            0.96969697
124   virginica   virginica        0.00000000             0.03030303            0.96969697
126   virginica   virginica        0.00000000             0.03030303            0.96969697
135   virginica  versicolor        0.00000000             0.92500000            0.07500000
138   virginica   virginica        0.00000000             0.03030303            0.96969697
141   virginica   virginica        0.00000000             0.03030303            0.96969697
143   virginica   virginica        0.00000000             0.03030303            0.96969697
144   virginica   virginica        0.00000000             0.03030303            0.96969697
>
```

Fonte: Adaptado pelo autor

2.2.2 Associação

Essa tarefa geralmente é utilizada em grandes conjuntos de dados para identificar a associação entre itens. Conforme explica Tan, Steinbach e Kumar (2006), devido ao grande acúmulo de dados referentes a operações diárias, como por exemplo registro de compras de clientes em grandes redes de supermercados, é possível através desta tarefa aprender sobre comportamento de compra de seus clientes.

Coleman e Ahlemeyer-Stubbe (2014) se referem a esta tarefa como análise de cesta de mercado. Eles afirmam que é possível, além de identificar quais produtos são comprados juntos, verificar a possibilidade de clientes comprarem em diferentes empresas ou ainda, a probabilidade de adquirirem certos serviços juntos.

Um dos algoritmos mais famosos desta etapa é o Apriori, conforme destaca Passos e Goldschmidt (2005). Muitos outros foram baseados neste como GSP, DHP, Partition, DIC, Eclat, MaxEclat, Clique e MaxClique. Como todos estes têm em sua base de funcionamento o Apriori, basicamente há duas etapas no seu funcionamento: encontrar todos os conjuntos de itens frequentes para então, a partir do conjunto de itens frequentes, gerar as regras de associação.

2.2.2.1 Exemplo de Associação

Para exemplificar o funcionamento da tarefa de associação, usou-se a linguagem R e o *dataset* Groceries, que faz parte da biblioteca de exemplos contidas no pacote de instalação do R.

Esta base de dados, Groceries, contém 9.835 linhas onde cada linha representa uma transação de mercado. Pode-se pensar em uma transação como um cupom fiscal que recebemos ao efetuar uma compra, que por exemplo, pode conter 3 itens como pão, manteiga e leite. A mesma lógica aplica-se para esta base de dados, cada linha representa um cupom fiscal e cada coluna representa um item deste cupom.

A figura 13 representa um resumo desta base de dados utilizando-se a função *summary* em R. Pode-se observar que há 9835 linhas contendo 169 colunas. Os três itens mais frequentes são *whole milk*, *other vegetables* e *rolls/buns*.

Na parte inferior da figura 13 há três exemplos de linhas contidas na base, cada uma contem três itens, porém isso não indica que todas as linhas da base se restringem a três itens por linhas, algumas linhas podem conter mais e outras menos itens.

Figura 13 - Resumo *dataset* Groceries

```
> summary(Groceries)
transactions as itemMatrix in sparse format with
 9835 rows (elements/itemsets/transactions) and
 169 columns (items) and a density of 0.02609146

most frequent items:
      whole milk other vegetables       rolls/buns             soda           yogurt          (Other)
            2513             1903             1809             1715             1372            34055

element (itemset/transaction) length distribution:
sizes
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20   21
2159 1643 1299 1005  855  645  545  438  350  246  182  117   78   77   55   46   29   14   14    9   11
  22   23   24   26   27   28   29   32
   4    6    1    1    1    1    3    1

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   4.409   6.000  32.000

includes extended item information - examples:
       labels  level2           level1
1 frankfurter sausage meat and sausage
2     sausage sausage meat and sausage
3  liver loaf sausage meat and sausage
```

Fonte: Adaptado pelo autor

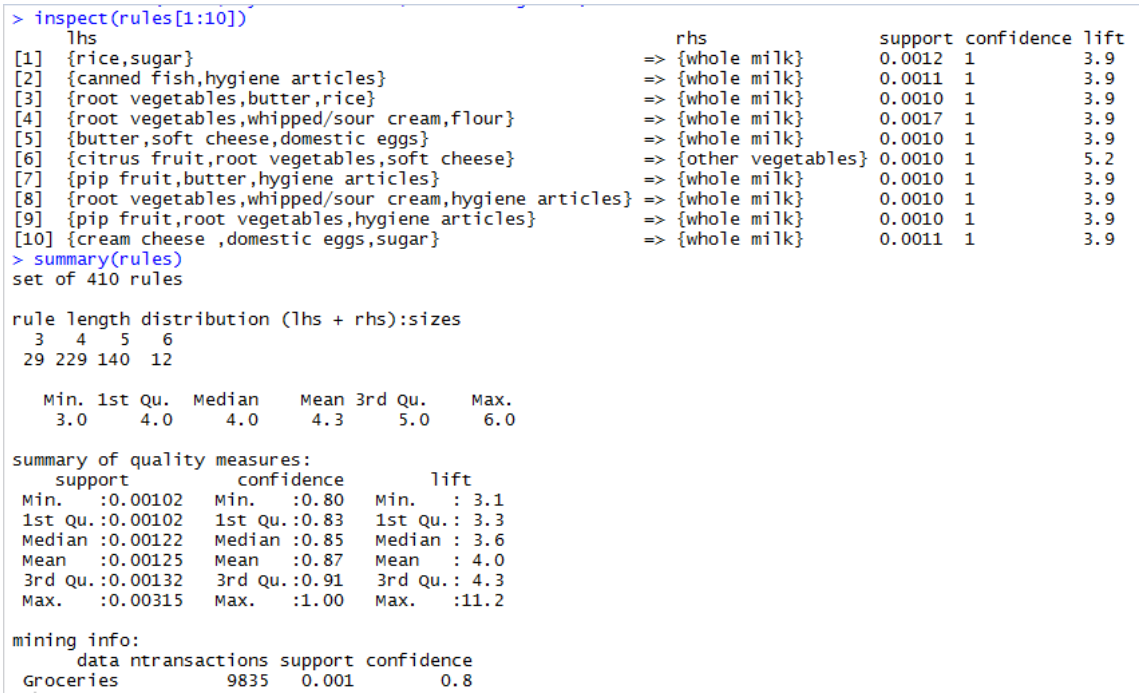
Antes de começar a utilizar a função *apriori*, presente na biblioteca *arules*, é importante definir dois parâmetros necessários para sua execução:

- *Support (supp)*: Suporte é uma indicação da frequência com que o conjunto de itens aparece no banco de dados, por exemplo o conjunto x = (leite, cereal) tem suporte de 1/5 = 0.20, ou seja, este ocorre em 20% de todas as transações.
- *Confidence (conf)*: O valor de confiança de uma regra, X => Y, dentro de um conjunto de transações T, é a proporção das transações que contém X que também contém Y. Por exemplo, a confiança da regra {leite, fralda} => {cerveja} é obtida dividindo a contagem de suporte para (leite, fraldas, cerveja), que aparecem apenas 2 vezes em uma mesma transação, pela contagem de suporte para (leite, fraldas). Uma vez que há 3 transações que contêm leite e fraldas, a confiança para esta regra é 2/3 = 0,67.

Neste exemplo será utilizado como valores para os parâmetros mencionados supp = 0.001, conf = 0.8, assim o algoritmo retornará todas as regras que estão presentes em pelo menos uma transação e que também sejam confiáveis em 80% das vezes ou mais.

A Figura 14 mostra exemplo das 10 primeiras regras geradas através da execução do algoritmo, ordenadas em ordem decrescente pela coluna de *confidence.* Essa figura também mostra outros dados referentes a massa de regras geradas como número de regras e quantidade de regras separada por número de itens.

Figura 14 - Resumo do conjunto de regras geradas para o *dataset* Groceries



```
> inspect(rules[1:10])
     lhs                                                  rhs                support confidence lift
[1]  {rice,sugar}                                      => {whole milk}       0.0012  1          3.9
[2]  {canned fish,hygiene articles}                    => {whole milk}       0.0011  1          3.9
[3]  {root vegetables,butter,rice}                     => {whole milk}       0.0010  1          3.9
[4]  {root vegetables,whipped/sour cream,flour}        => {whole milk}       0.0017  1          3.9
[5]  {butter,soft cheese,domestic eggs}                => {whole milk}       0.0010  1          3.9
[6]  {citrus fruit,root vegetables,soft cheese}        => {other vegetables} 0.0010  1          5.2
[7]  {pip fruit,butter,hygiene articles}               => {whole milk}       0.0010  1          3.9
[8]  {root vegetables,whipped/sour cream,hygiene articles} => {whole milk}   0.0010  1          3.9
[9]  {pip fruit,root vegetables,hygiene articles}      => {whole milk}       0.0010  1          3.9
[10] {cream cheese ,domestic eggs,sugar}               => {whole milk}       0.0011  1          3.9
> summary(rules)
set of 410 rules

rule length distribution (lhs + rhs):sizes
  3   4   5   6
 29 229 140  12

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    3.0     4.0     4.0     4.3     5.0     6.0

summary of quality measures:
    support          confidence        lift
 Min.   :0.00102   Min.   :0.80   Min.   : 3.1
 1st Qu.:0.00102   1st Qu.:0.83   1st Qu.: 3.3
 Median :0.00122   Median :0.85   Median : 3.6
 Mean   :0.00125   Mean   :0.87   Mean   : 4.0
 3rd Qu.:0.00132   3rd Qu.:0.91   3rd Qu.: 4.3
 Max.   :0.00315   Max.   :1.00   Max.   :11.2

mining info:
      data ntransactions support confidence
 Groceries          9835   0.001        0.8
```

Fonte: Adaptado pelo autor

Como visto no exemplo, a tarefa de associação se encaixa perfeitamente na busca para identificar itens que possam estar associados. Para isso gera-se regras que podem ser interpretadas como “se isso, então aquilo”. No caso das transações mencionadas, essa tarefa gerou 410 regras, onde é possível verificar na massa de dados quais os itens que estão mais suscetíveis de serem comprados em conjunto. Por este ser o uso mais comum desta tarefa ela também recebe o nome de *Market Basket Analysis.*

2.2.3 Regressão

Conceitos de regressão foram publicados pela primeira vez no início de 1800, antes mesmo da era dos computadores, e representaram um grande avanço no campo da estatística. Porém ainda hoje há boas razões para estudá-los conforme aponta Hastie, Tibshirani e Friedman (2009).

“A regressão é uma técnica de modelagem preditiva onde a variável alvo a ser avaliada é continua. ” (TAN; STEINBACH; KUMAR, 2006) Exemplos do uso dessa tarefa incluem a previsão da quantidade de precipitação em uma região baseada nas características dos ventos, previsão de índice na bolsa de valores utilizando outros indicadores econômicos ou até mesmo estimativa da

probabilidade de um paciente sobreviver, dado o resultado de um conjunto de diagnósticos de exames.

De acordo com Coleman e Ahlemeyer-Stubbe (2014), regressão linear é um dos métodos mais fáceis de ser entendido e aplicado, podendo gerar previsões muito exatas e precisas. Para derivar um modelo de regressão linear simples, em que existe apenas uma variável de previsão, os dados são representados como pontos num modelo de duas dimensões, em que o eixo Y representa o alvo e o eixo X acomoda as variáveis preditoras. Após colocar todos os pontos no modelo, uma linha de regressão linear é traçada (linha reta) entre os pontos, de tal maneira que a distância entre a linha e todos os pontos é a menor possível.

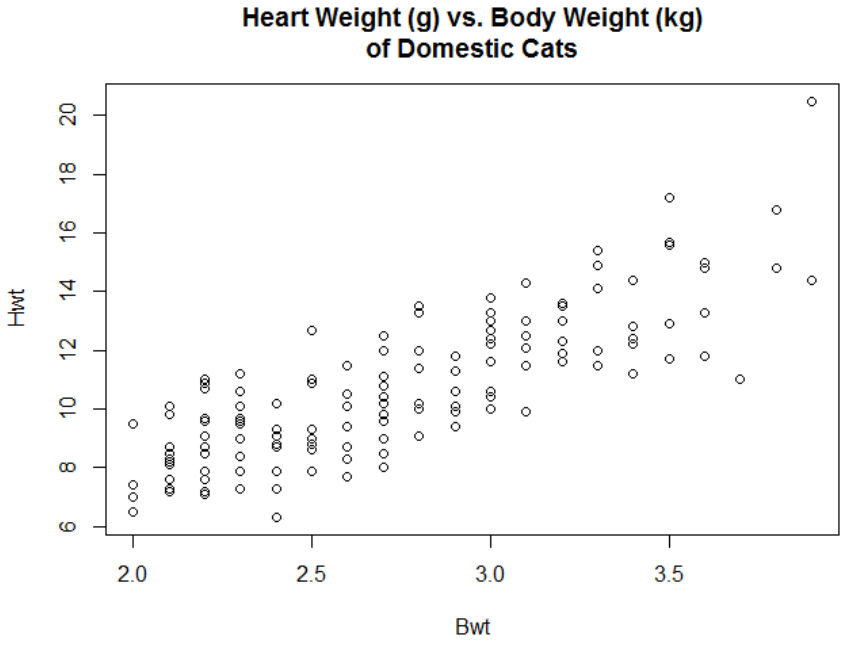
2.2.3.1 Exemplo de Regressão

Para exemplificar o funcionamento da tarefa de regressão, usou-se a linguagem R e o *dataset cats*, que faz parte da biblioteca de exemplos contidas no pacote de instalação do R.

Essa base de dados contém 144 observações contendo 3 variáveis: massa corporal em kg, peso do coração em gramas e sexo de 144 gatos. A tarefa de regressão foi utilizada para buscar a relação entre a massa corporal de cada gato com o tamanho do coração.

A Figura 15 mostra a distribuição das variáveis de bwt = massa corporal x hwt = tamanho do coração, para todas as 144 observações. Nesse exemplo temos a variável de tamanho do coração no eixo Y, que representa o alvo e massa corporal no eixo X.

Figura 15 - Distribuição de massa corporal x tamanho do coração



Fonte: Adaptado pelo autor

É possível observar uma relação razoavelmente linear entre as variáveis. Utilizando a linguagem R e a formula de correlação (with(cats, cor.test(Bwt, Hwt))) verifica-se que a correlação é positiva, 0.8041274, ou seja, à medida que a massa corporal do gato aumenta, o tamanho do seu coração também aumenta.

Após verificado visualmente a distribuição das variáveis e a correlação, utilizou-se a função *lm (linear model)* para criar o modelo de regressão, conforme mostra a Figura 16.

Figura 16 - Criação do modelo de regressão

```
> lm(Hwt ~ Bwt, data=cats)

Call:
lm(formula = Hwt ~ Bwt, data = cats)

Coefficients:
(Intercept)          Bwt
    -0.3567       4.0341

>
```

```
> summary(lm.out)

Call:
lm(formula = Hwt ~ Bwt, data = cats)

Residuals:
    Min      1Q  Median      3Q     Max
-3.5694 -0.9634 -0.0921  1.0426  5.1238

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.3567     0.6923  -0.515    0.607
Bwt           4.0341     0.2503  16.119   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.452 on 142 degrees of freedom
Multiple R-squared:  0.6466,    Adjusted R-squared:  0.6441
F-statistic: 259.8 on 1 and 142 DF,  p-value: < 2.2e-16
```

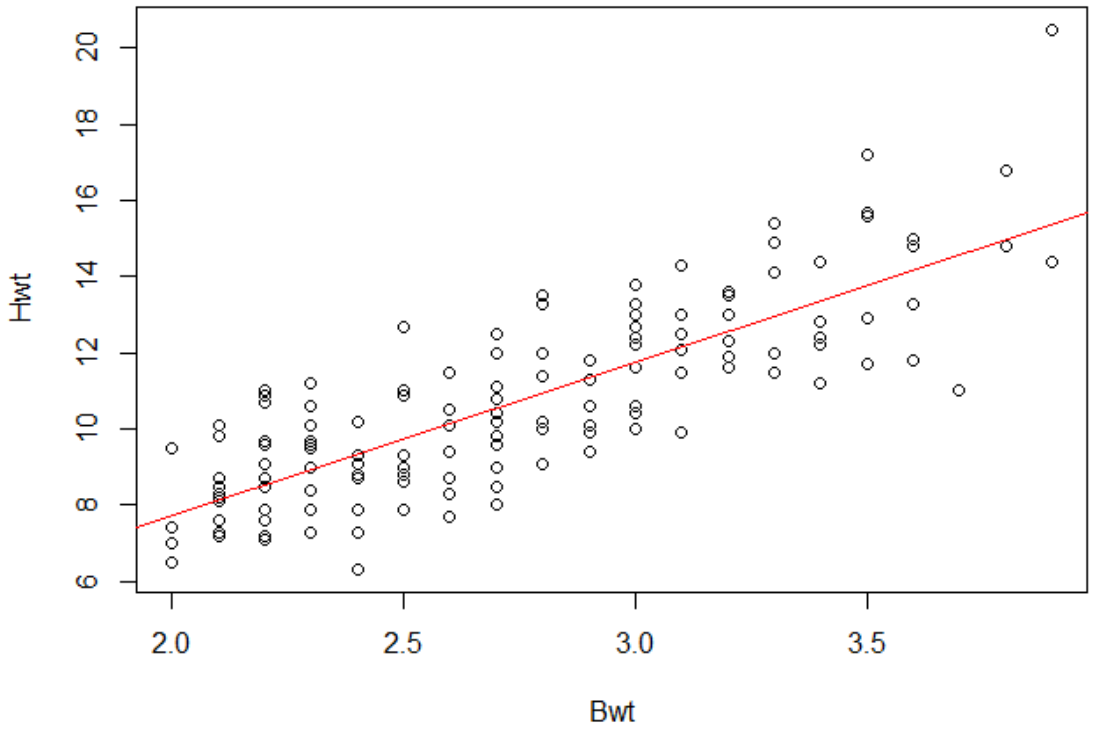
Fonte: Adaptado pelo autor

Portanto, a equação de regressão completa é tamanho do coração que se quer encontrar (variável do eixo Y) = -0,3567 + 4,0341 * massa corporal do

gato (variável do eixo X). Por exemplo, se um gato possui massa corporal igual a 4 pode-se tentar prever o tamanho do coração utilizando a equação -0,3567 + 4,0341 * 4 que resultaria em um coração de 15,77 gramas.

Observando a Figura 17, onde desenhou-se a reta da equação sobre a distribuição das variáveis, podemos verificar que o número descoberto na equação faz parte desta reta.

Figura 17 - Reta do modelo linear



Fonte: Adaptado pelo autor

Conforme mostrou-se no exemplo, um modelo de regressão pode ser uma maneira muito útil e de fácil utilização para predição de valores. Essa técnica tem grande uso no mercado de ações, onde através da relação entre variáveis, como por exemplo, alta de uma determinada moeda em relação à inflação de um país buscando assim prever as possíveis variações de valor desta mesma moeda.

2.2.4 Análise de Grupos

A análise de grupos também conhecida como *cluster analysis* ou segmentação de dados, tem uma variedade de objetivos conforme aponta Hastie, Tibshirani e Friedman (2009). Todos estes objetivos estão relacionados a segmentação ou agrupamento de uma coleção de objetos em subconjuntos,

clusters, de tal forma que os objetos contidos dentro de cada agrupamento são mais relacionados entre si do que objetos contidos em diferentes agrupamentos.

"*Clustering* é o processo de organizar objetos em grupos semelhantes, descobrindo os limites entre estes grupos algoritmicamente, usando um número diferente de algoritmos e métodos estatísticos." (DEAN, 2014, p.132, tradução nossa) Análise de *cluster* não faz qualquer distinção entre variáveis dependentes e independentes. Ele examina todo o conjunto de dados para descobrir as relações de semelhança entre os objetos, a fim de identificar os *clusters*.

Passos e Goldschmidt (2005), ainda destacam que diferente da classificação, que tem rótulos pré-definidos, a clusterização precisa identificar os rótulos ou grupos de maneira automática. Por esta razão esse método é também denominado de indução não supervisionada.

Os procedimentos de análise de grupos não têm mecanismos de seleção para as variáveis de entrada, de modo que o analista deve decidir quais variáveis são usadas com a ajuda de conhecimento do negócio e pré-análises feitas em etapas anteriores. Existem vários métodos de análise de agrupamento, que diferem na maneira como os grupos são construídos, o tipo de variáveis de entrada que pode ser utilizado e a velocidade de encontrar os *clusters*. (COLEMAN; AHLEMEYER-STUBBE, 2014). Os métodos mais comuns são análise de agrupamento hierárquico, K-Means e DBSCAN. (TAN; STEINBACH; KUMAR, 2006)

2.2.4.1 Exemplo de grupos

Para exemplificar o funcionamento da tarefa de análise de grupos, usou-se a linguagem R e o *dataset iris*, que faz parte da biblioteca de exemplos contidas no pacote de instalação do R.

Esse *dataset* é composto por 150 observações contendo 5 variáveis: comprimento e largura da sépala, comprimento e largura das pétalas e espécie de íris. Antes de demonstrar o uso dessa tarefa, removeu-se a variável espécie para então ser utilizada a tarefa de análise de grupos e determinar o grupo, ou espécie, a qual cada observação pertence levando em conta as 4 variáveis

restantes na base de dados. A Figura 18 mostra a composição do *dataset* iris2, que é a adaptação criada para demonstrar este exemplo.

Figura 18 - Composição *dataset* iris2

```
> summary(iris2)
  Sepal.Length    Sepal.Width     Petal.Length    Petal.Width
 Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100
 1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300
 Median :5.800   Median :3.000   Median :4.350   Median :1.300
 Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
 3rd Qu.:6.400   3rd Qu.:3.300   3rd Qu.:5.100   3rd Qu.:1.800
 Max.   :7.900   Max.   :4.400   Max.   :6.900   Max.   :2.500
>
```

Fonte: Adaptado pelo autor

Para este exemplo foi usado o algoritmo *K-means*, já presente na instalação padrão da linguagem R. *K Means Clustering* é um algoritmo de aprendizagem não supervisionado que tenta agrupar dados com base na sua similaridade. A aprendizagem não supervisionada significa que não há nenhum resultado a ser previsto, diferente da tarefa de regressão, e o algoritmo apenas tenta encontrar padrões nos dados. Em *K-means*, temos que especificar o número de *clusters* que queremos que os dados sejam agrupados. O algoritmo atribui aleatoriamente cada observação a um cluster e localiza o centroide[2] de cada *cluster*. Em seguida, o algoritmo roda em duas interações: Reatribuir pontos de dados para o *cluster* cujo centroide é o mais próximo. Calcular o novo centroide de cada *cluster*.

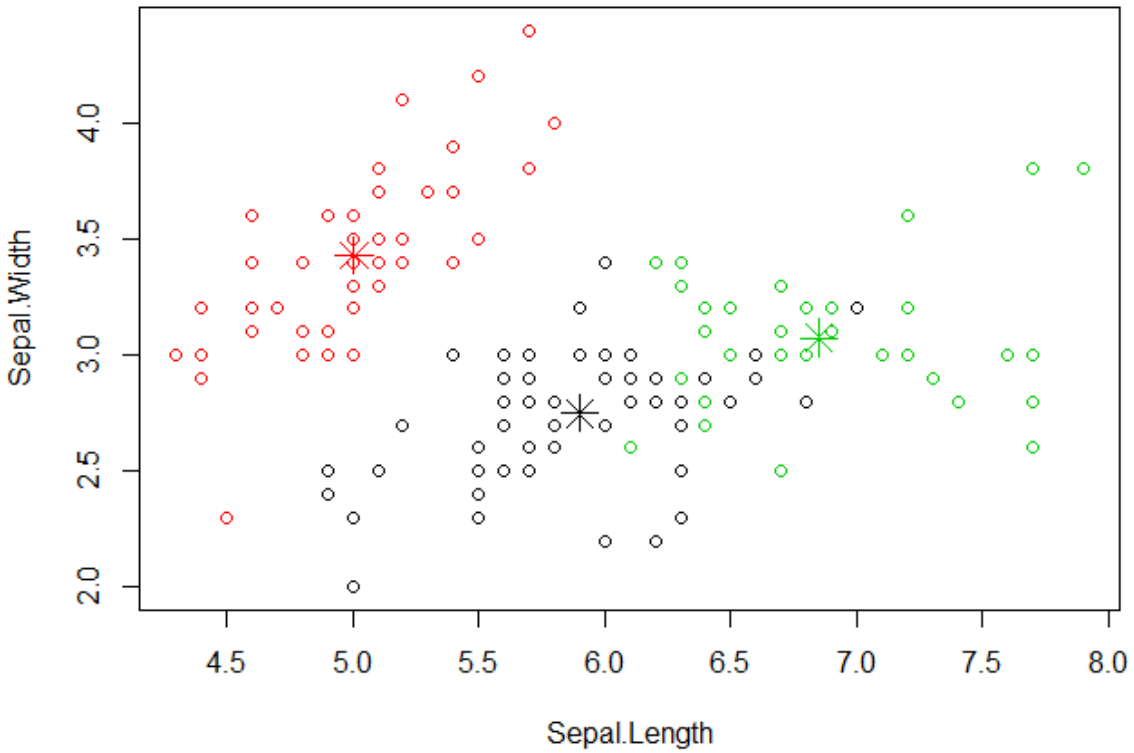
Após a execução do algoritmo, cria-se três *clusters*, kmeans(iris2, 3), pode-se verificar a divisão dos grupos e seus centroides na Figura 19. Cada *cluster* é identificado por uma cor e seus centroides são indicados através de asteriscos.

[2] O centroide do *cluster* é o meio de um *cluster*. Um centroide é um vetor contendo um número para cada variável, onde cada número é a média de uma variável para as observações nesse conjunto.

Figura 19 - Comparação dos grupos originais, resultado *K-means* e divisão dos grupos

```
> table(iris$Species, kmeans.result$cluster)

            1  2  3
setosa      0 50  0
versicolor 48  0  2
virginica  14  0 36
```



Fonte: Adaptado pelo autor

Na tabela apresentada na Figura 19, podemos ver na comparação entre os 3 clusters criados na execução do algoritmo e os grupos originais separados por espécies que houve uma divergência nas quantidades de itens por grupo. O *dataset* original era formado por 150 observações separadas em 3 espécies, cada grupo de espécie era formado por 50 observações. Porém, nos novos grupos formados agrupando itens similares através das variáveis de comprimento e largura da sépala, comprimento e largura das pétalas, obteve-se 62 itens no primeiro grupo, 50 no segundo e 38 no terceiro.

Conforme visto no exemplo, a tarefa de análise de grupos é muito usada para analisar e agrupar itens de acordo com suas similaridades, podendo incluir mais ou menos características de acordo com a necessidade do estudo.

2.3 Considerações Finais

Como destacado neste capítulo, a mineração de dados é uma importante etapa na descoberta de conhecimento, que é largamente utilizada tanto no meio corporativo quanto acadêmico. No item 2.1, segundo parágrafo, o autor Dean (2014) mostra como pesquisas que utilizam *Data Mining* podem mudar a vida de muitas pessoas.

Os autores Katoua (2013), Fayyad (1996) e Goldschmidt (2005) apresentam etapas muito parecidas para obtenção dos resultados de KDD. Destaca-se o processo de obter melhor conhecimento dos dados que estão em análise como fundamental para traçar os objetivos da mineração, bem como necessário na escolha das tarefas e técnicas a serem utilizadas. Para então, em um último momento se estudar os conhecimentos obtidos. Essas etapas são mencionadas pelos três autores, e merecem destaque por serem essenciais em qualquer processo de descoberta de conhecimento.

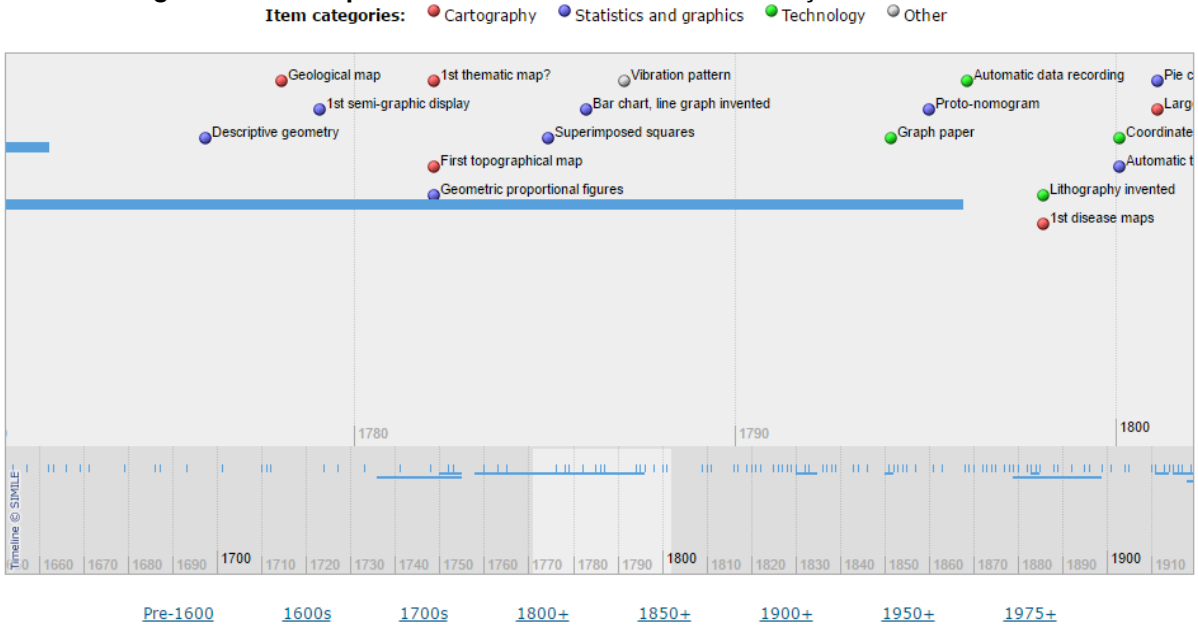
No que tange as tarefas de mineração de dados, há muitas opções já conhecidas. Porém, baseado no conjunto de dados que será utilizado para elaboração deste trabalho, as 4 tarefas aqui apresentadas: classificação, associação, regressão e análise de grupos, são as que demonstram maior adequação à natureza dos dados.

3. VISUALIZAÇÃO DOS DADOS

A visualização de dados é uma área ativa de aplicação e pesquisa. Apesar de ser um tema atual largamente discutido, técnicas de visualização já estão presentes há muito tempo na história da humanidade. Conforme mostra o projeto *Milestones in the History of Thematic Cartography, Statical Graphics, and Data Visualization,* produzido pelos professores Michel Friendly e Daniel J. Denis (2001), no século XVI, técnicas e instrumentos para observação e medição de grandezas físicas precisas já haviam sido desenvolvidos.

Em seu portal na internet, os autores do projeto apresentam um gráfico interativo, onde é possível conferir os principais *milestones* da história da visualização gráfica. Conforme mostra a Figura 20, pode-se observar a criação do gráfico de barras por volta de 1780.

Figura 20 - Principais Marcos na História da visualização de dados.

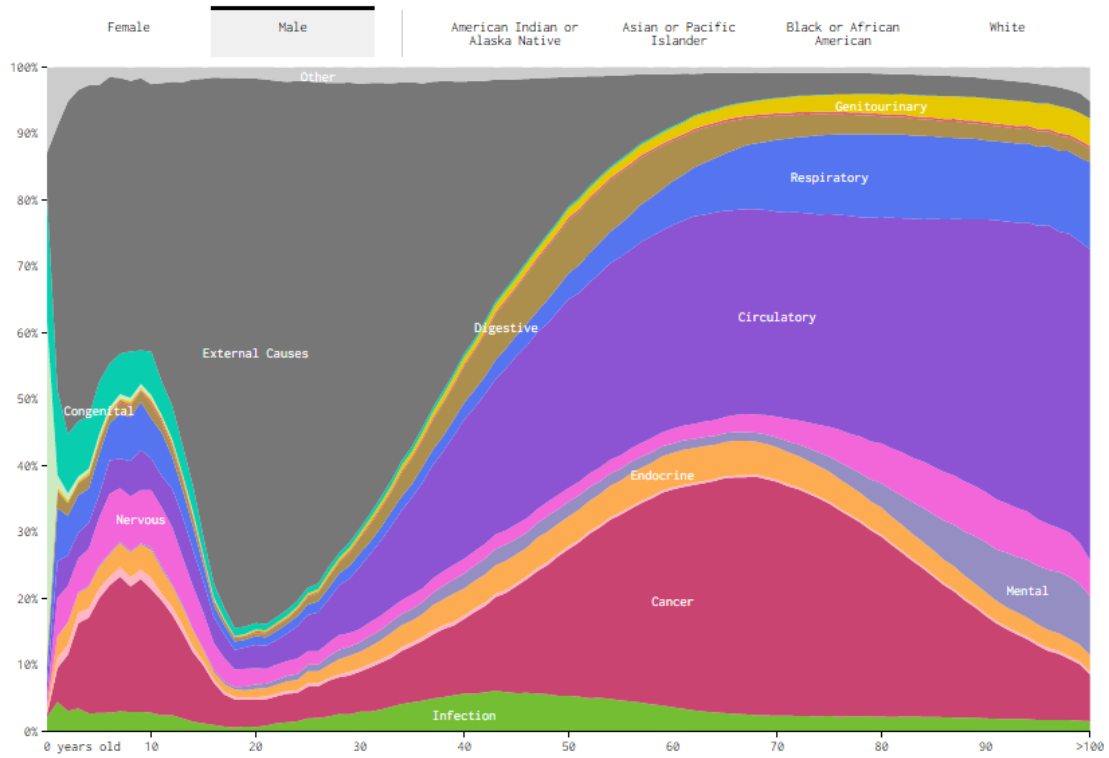


Fonte: http://www.datavis.ca/milestones/

Desde os primeiros indícios de uso de técnicas de visualização até os dias atuais, o termo “visualização de dados” vai além de apenas métodos tradicionais de visualizar informação como gráficos de pizza ou de barras. Yuk e Diamond (2014, p.7, tradução nossa) descrevem como “[...]o estudo de como representar dados usando uma aproximação visual ou artística em vez de métodos de relatórios tradicionais. ”

Uma amostra desta evolução e de como a visualização de dados pode ser utilizada para apresentação das mais variadas informações é o site chartporn.org, que reúne trabalhos dos mais variados autores e outros sites, todos ligados à área de visualizações. Ele exibe uma variedade de técnicas de representação de dados utilizando humor, criatividade e apresentando informações atuais em gráficos e visualizações simples, mas efetivos, como é o caso das Figuras 21 e 22, extraídas do portal:

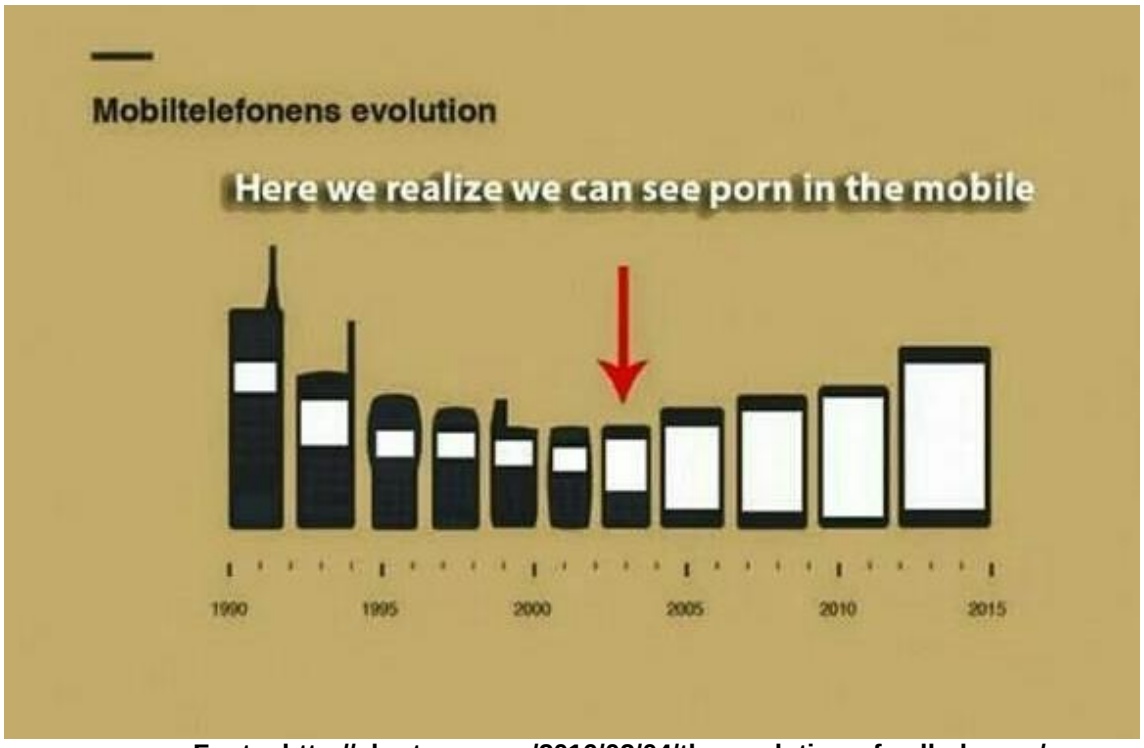
Figura 21 - Causas de morte de homens, por faixa etária, entre 2005 e 2014 no mundo.



Fonte: http://flowingdata.com/2016/01/05/causes-of-death/

A figura 21, mostra de forma arrojada e muito eficiente as causas de mortes entre homens nos anos entre 2005 e 2014 com idades entre 0 e 100 anos. De forma interativa, ainda seria possível visualizar os mesmos dados entre mulheres ou etnia. Facilmente é possível constatar que entre 50 e 80 anos as principais causas de morte são o câncer e problemas circulatórios.

Figura 22 - Evolução dos Celulares.



Fonte: http://chartporn.org/2016/02/04/the-evolution-of-cell-phones/

A figura 22 foi extraída do mesmo portal. De forma bem humorada e apresentando um gráfico contendo formatos de aparelhos celulares, podemos ver uma forte tendência de diminuição do tamanho dos aparelhos até pouco depois dos anos 2000. Porém, como aponta a Figura, por volta de 2003 com a popularização da internet e com a melhora na exibição de imagens através das telas dos aparelhos, essa tendência se inverteu e caminha para modelos maiores novamente, mas agora com telas extremamente grandes que ocupam quase toda área frontal dos aparelhos.

> "Visualização de dados (também conhecido como DataViz) é um tema quente. Com o desenvolvimento de mais fontes de dados, tais como plataformas de mídia social, fotos e *reviews* de clientes, *Big Data* tornou-se uma preocupação para pequenas empresas e grandes corporações. Dados estão vindo de todas partes do negócio como finanças, atendimento ao cliente e vendas, usá-los de forma eficaz ajuda você a ganhar uma vantagem competitiva. "(Yuk; Diamond, 2014, p.1, tradução nossa)

Atualmente o termo visualização de dados está fortemente atrelado ao *Big Data*, assim como mostrado no item 2.1, onde o autor Jean Dean (2014) considera que com o início da era do *Big Data* empresas e pesquisadores começaram a prestar mais atenção no valor futuro dos dados. Porém, viu-se que havia a necessidade de interpretar e visualizar todo o conhecimento acumulado. Matt Asay (2016) escreveu no portal InfoWorld na sua matéria intitulada *Data*

visualization: Showing isn't always telling que a visualização de dados é uma das áreas mais quentes do *Big Data* e pode revelar percepções dos dados que até então estariam ocultas para analistas. Ele ainda afirma que a visualização dos dados de maneira eficiente pode complementar a intuição humana.

A visualização de dados pode ser uma importante aliada no processo de descoberta de conhecimento. Conforme demonstra Katoua (2013) no item 2.1 deste trabalho, a última etapa do processo é a de interpretação dos dados, que através de técnicas de visualizações pode acontecer de maneira mais eficiente. "Sistemas de visualização baseados em computadores fornecem representações visuais de *datasets* destinados a ajudar as pessoas a realizar tarefas de forma mais eficaz. " (Munzer, 2014, p.1, tradução nossa).

"Gráficos fornecem uma excelente abordagem para explorar dados e são essenciais para a apresentação dos resultados. " (Chen; Härdle; Unwin, 2008, p.4, tradução nossa).

3.1 Técnicas de Visualização de dados

O campo de estudo e uso de visualização dos dados vai além de gráficos estáticos. Ele abrange imagens e com a ajuda de computadores também se utiliza de recursos interativos para agregar mais valor ao conteúdo que está sendo apresentado.

Chiasson e Gregory (2016) comparam as etapas de preparação e visualização dos dados ao ato de cozinhar. Primeiro juntam-se todos os ingredientes necessários, limpa-se e aplica-se alguma técnica a cada ingrediente, como descascar, esmagar, separar ou cozinhar, antes de combiná-los. O mesmo acontece em etapas do processo de KDD. Ao combinar os ingredientes para se preparar um prato, alguns destes serão destacados enquanto outros farão o papel de suporte, já outros vão desaparecer. Porém, não basta apenas prestar atenção na harmonização dos ingredientes, é necessário um bom visual para que as características que se deseja ressaltar no prato fiquem em evidência.

Chiasson e Gregory (2016) ainda utilizam o exemplo de uma sopa vegetariana, pois para conseguir destaque dos vegetais utilizados, uma boa prática seria cortá-los, ao invés de esmagá-los. Similarmente, isso acontece

quando se decide quais tipos de gráficos serão usados para montar a visualização de um determinado conjunto de dados. Fatores como tipo de dados e a informação ou mensagem que se deseja enfatizar influenciam nos gráficos a serem utilizados.

Chen, Härdle e Unwin (2008) afirmam que mesmo com o uso de gráficos por um longo tempo, não há um corpo de conhecimento consistente sobre o assunto. Mesmo assim, eles defendem a ideia de separar técnicas de visualização de acordo com o seu propósito, apresentação ou exploração de dados.

Munzer (2014) apresenta um *framework* de três passos para análise e construção de gráficos:

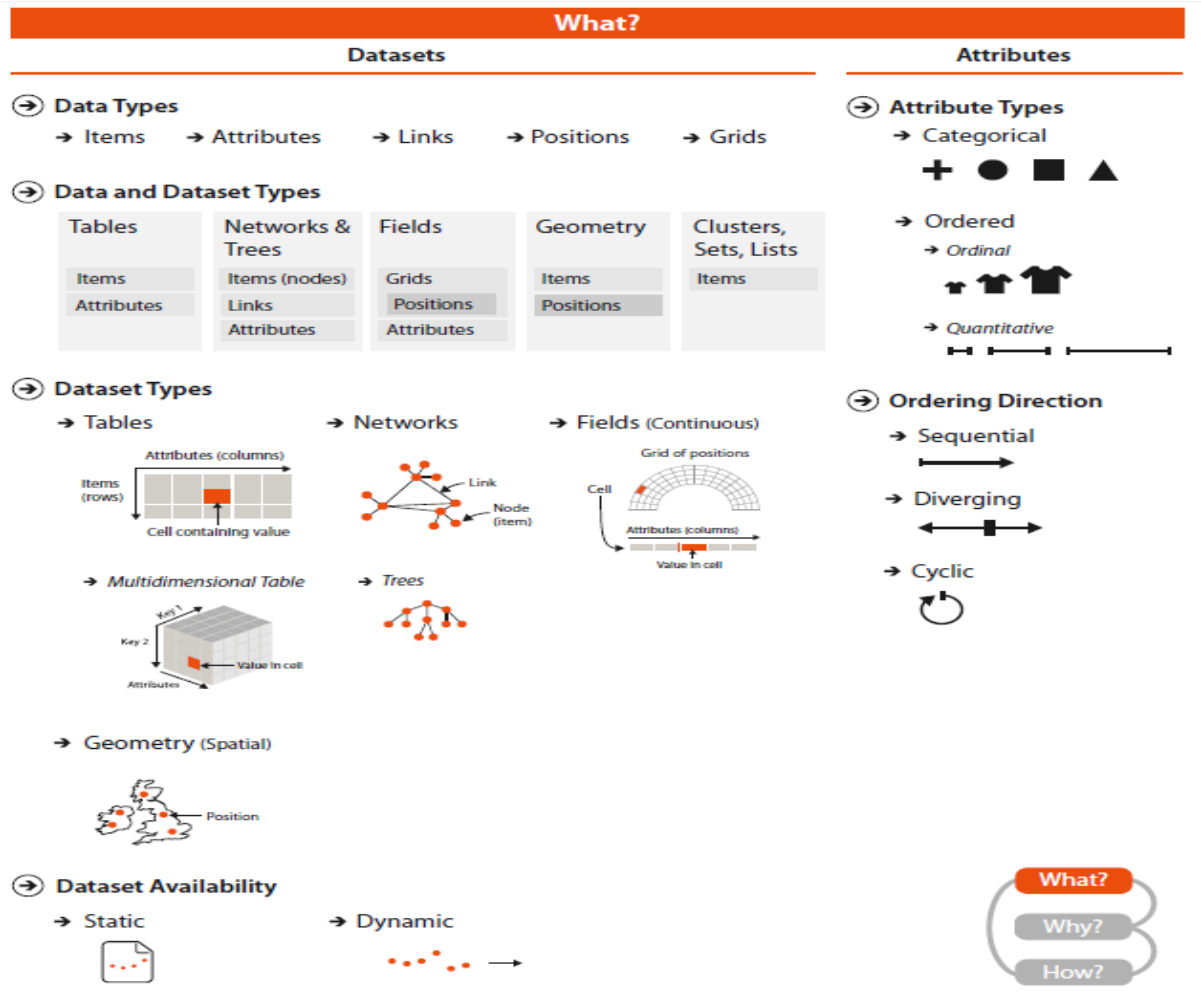
O primeiro, como mostra a Figura 23, seria a identificação do que se pode ser visualizado. Ela define tabelas, *networks*, campos ou células e geometria (espacial) como sendo os quatro tipos de *datasets* básicos. Porém, ainda há outros tipos como *clusters*, conjuntos e listas. Estes tipos de *datasets* são formados de diferentes tipos de dados e os atributos destes dados ainda podem ser divididos em categóricos ou ordenados, ordinais e quantitativos.

O segundo passo do *framework*, se refere ao porquê de uma ferramenta de visualização estar sendo utilizada, conforme a Figura 24. O autor divide o porquê em dois grandes grupos:

- Ações: Trata-se de usar uma ferramenta de visualização para produzir ou consumir informações. Como explica o autor, consumir os dados para uma apresentação, descoberta ou apenas para alguma finalidade de uso próprio. Enquanto a descoberta de dados pode envolver geração ou verificação de uma hipótese.
- Objetivos: Para todos os tipos de dados em que se deseja descobrir tendências e valores discrepantes. Para um atributo, o objetivo pode ser um valor, o extremo dos valores mínimos ou máximos, a distribuição de todos os valores para determinado atributo. Para vários atributos o objetivo pode ser verificar dependências, correlação, ou similaridades entre eles.

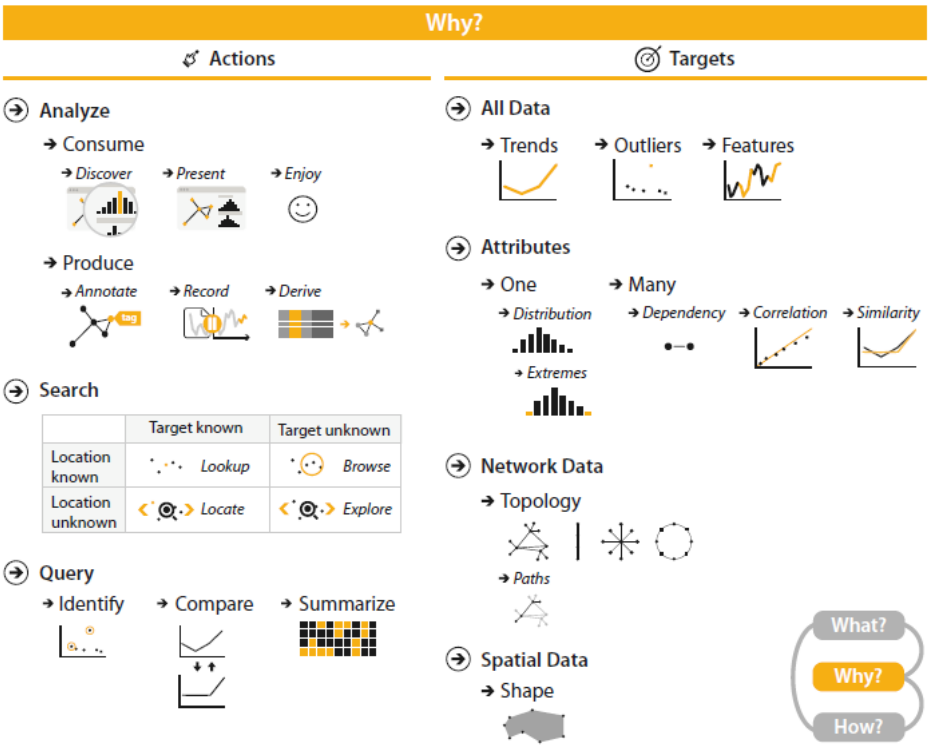
O passo final é ilustrado pela Figura 25, onde busca-se o como, ou seja, após o entendimento dos dados que serão analisados e o porquê, deve-se definir como a visualização será construída. Neste último passo define-se o design da visualização. Características como codificação dos dados dentro da visualização, forma de manipulação, exibição de múltiplas visualizações para determinado conjunto de dados e por fim redução de dados que o autor caracteriza como possibilidade de filtragem, agregação e embutir vários contextos em uma única visualização.

Figura 23 - 1º passo do framework de Munzer, What?



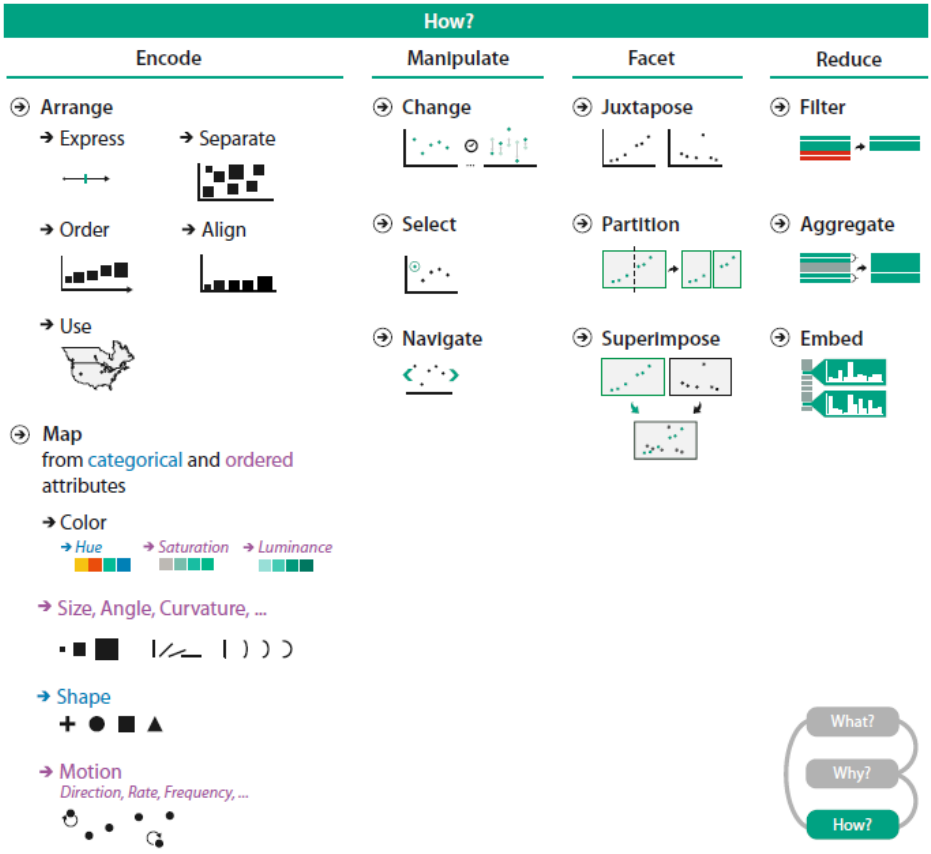
Fonte: Munzer 2014, p. 20

Figure 24 - 2º passo do framework de Munzer, Why?



Fonte: Munzer 2014, p. 42

Figura 25 - 3º passo do framework de Munzer, How?



Fonte: Munzer 2014, p. 58

SAS, *Statistical Analysis System* (Sistema de Análise Estatística) é uma grande empresa que fornece soluções de análise, com mais de 45 anos de experiência. Sua solução de mesmo nome, SAS, possui muitas ferramentas de visualização de grandes massas de dados. Ao verificar *white papers* da companhia sobre visualização de dados e criação de gráficos, são apresentadas algumas dicas para essa tarefa. Resumidamente destacam-se quatro etapas como básicas para tonar a visualização simples e eficiente:(SAS, 2016)

- Entender os dados que pretende-se visualizar, incluindo seu tamanho e cardinalidade;
- Determinar o que pretende-se visualizar e que tipo de informação será comunicada;
- Conhecer o público alvo desta visualização, entender como este público processa a informação visual;
- Usar um *design* que se encaixe com o que está sendo transmitido, de um jeito simples;

Chen, Härdle e Unwin (2008) não apresentam uma técnica ou receita de bolo para confecção de uma boa visualização, porém fazem a seguinte afirmação:

> "O que é plotado vem em primeiro lugar, e sem conteúdo nenhuma quantidade de *design* inteligente pode trazer significado para uma exibição. Um bom gráfico é sempre parte de um todo maior, o contexto, o que proporciona a sua relevância. Assim, um bom gráfico irá complementar outros materiais relacionados e se encaixa, tanto em termos de conteúdo e também no que diz respeito ao estilo e layout. Finalmente, se um gráfico é construído e bem desenhado, ele vai ficar bem." (Chen; Härdle; Unwin, 2008, p.58, tradução nossa)

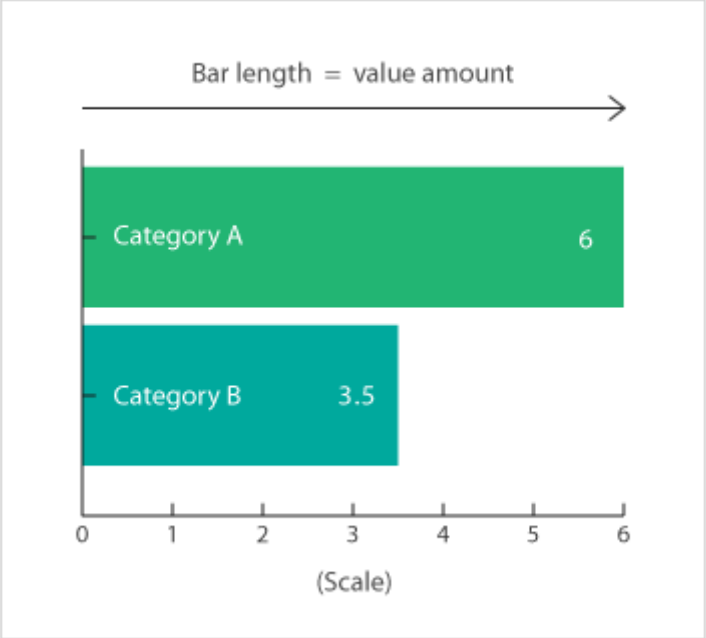
3.1.1 Gráfico de barras

O gráfico de barras é comumente usado para comparar quantidades de diferentes categorias ou grupos Chiasson e Gregory (2016) explica que gráfico de barras "[...]usa barras horizontais ou verticais, cujos comprimentos proporcionalmente representam valores em um conjunto de dados. Um gráfico com barras verticais também é chamado um gráfico de colunas ou *chart*".

Algumas vezes quando há muitas barras em um gráfico e estas são dispostas muito próximas umas das outras a empresa SAS (2016) ainda adiciona como boa pratica o uso de cores.

O portal datavizcatalogue.com (2016) ainda cita a diferença entre gráfico de barras e histogramas, que apesar de também ser constituído de colunas, este representa desenvolvimento contínuo em um intervalo de dados, já o gráfico de barras apresenta dados categóricos. A Figura 26 mostra um exemplo de gráfico de barras.

Figura 26 - Anatomia do gráfico de barras

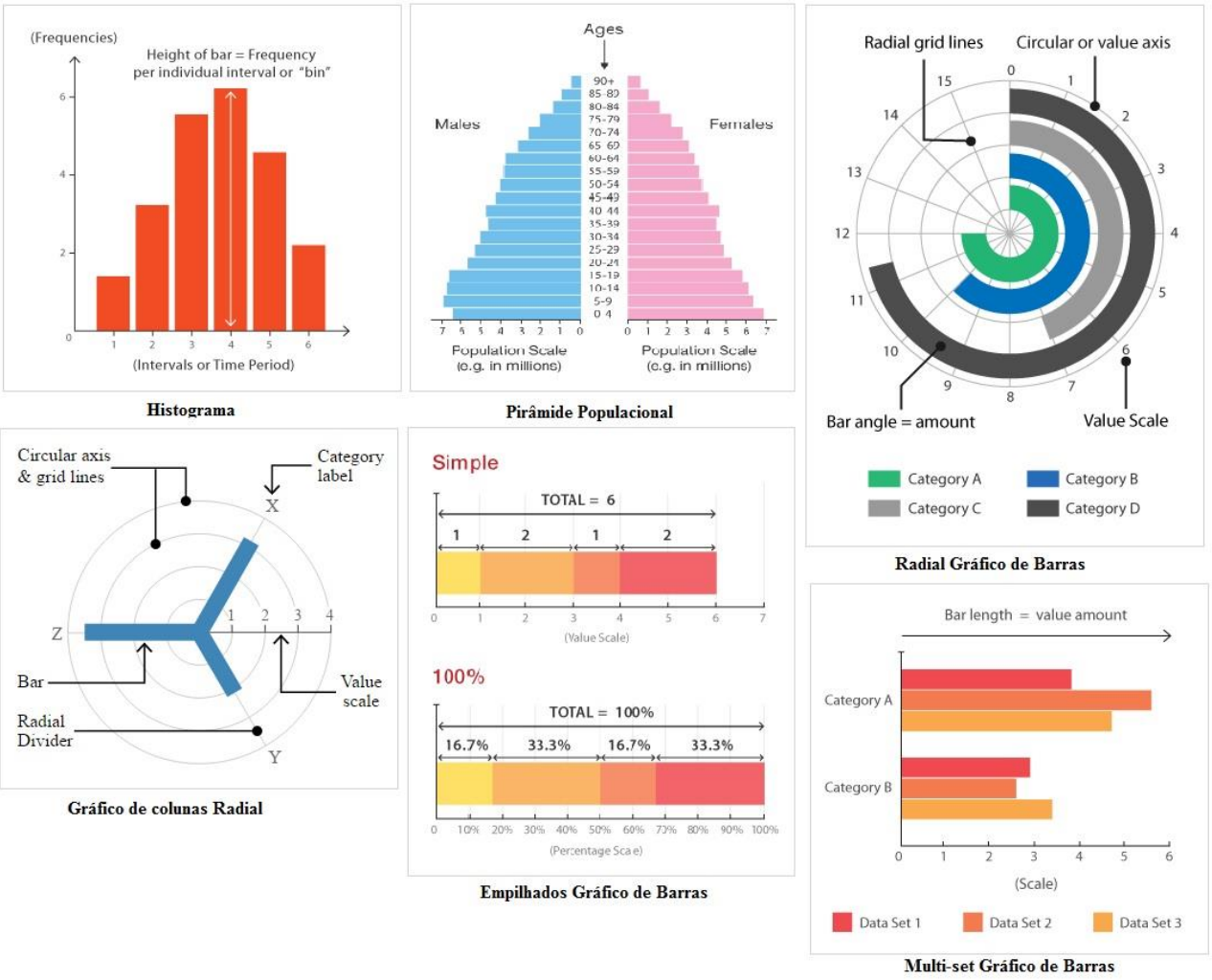


Fonte: http://www.datavizcatalogue.com/methods/bar_chart.html

3.1.1.1 Gráficos similares

Há variações do gráfico de barras conforme aponta o portal datavizcatalogue.com (2016) que podem ser conferidos na Figura 27.

Figura 27 - Tipos de Gráficos de Barra



Fonte: Adaptado pelo Autor

3.1.2 Gráfico de Pizza

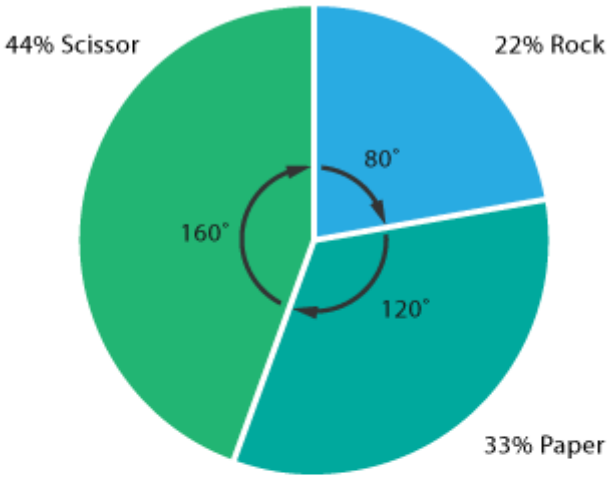
Gráfico de pizza ou também *piechart* é muito utilizado em apresentações e escritórios. Este tipo de gráfico ajuda a mostrar proporções e porcentagens entre as categorias. Cada "fatia" representa a proporção de cada categoria enquanto o círculo completo representa a soma total de todos os dados totalizando 100%. (datavizcatalogue.com, 2016)

Porém há um fator importante a ser considerado conforme mostra SAS (2016, p.6, tradução nossa):

> "[...] eles podem ser difíceis de interpretar porque o olho humano tem dificuldade em estimar a área e comparar os ângulos. Um outro desafio com o uso de um gráfico de pizza para análise é que é difícil comparar fatias do gráfico que são semelhantes em tamanho, mas não estão localizadas ao lado uma da outra."

Outros pontos a serem levados em consideração ao se usar um gráfico de pizza é a quantidade de valores que se pretende mostrar, pois quanto mais valores mais fatias serão acrescidas, o que pode tornar a visualização difícil. Apesar disso, comparar uma determinada categoria (uma fatia) dentro de um total em um único gráfico pode se mostrar muito eficiente. (datavizcatalogue.com, 2016) A Figura 28 mostra como fazer a divisão das fatias em um gráfico de pizza.

Figura 28 - Anatomia de um Gráfico de Pizza



Data			
Rock	Paper	Scissor	TOTAL
2	3	4	9
To calculate percentages			
2/9=22%	3/9=33%	4/9=44%	100%
Degrees for each "pie slice"			
(2/9) x 360 = 80°	(3/9) x 360 = 120°	(4/9) x 360 = 160°	360°

Fonte: http://www.datavizcatalogue.com/methods/pie_chart.html

3.1.2.1 Gráficos Similares

Há variações do gráfico de pizza conforme aponta o portal datavizcatalogue.com (2016) que podem ser conferidos na Figura 29.

Figura 29 - Tipos de Gráfico de Pizza

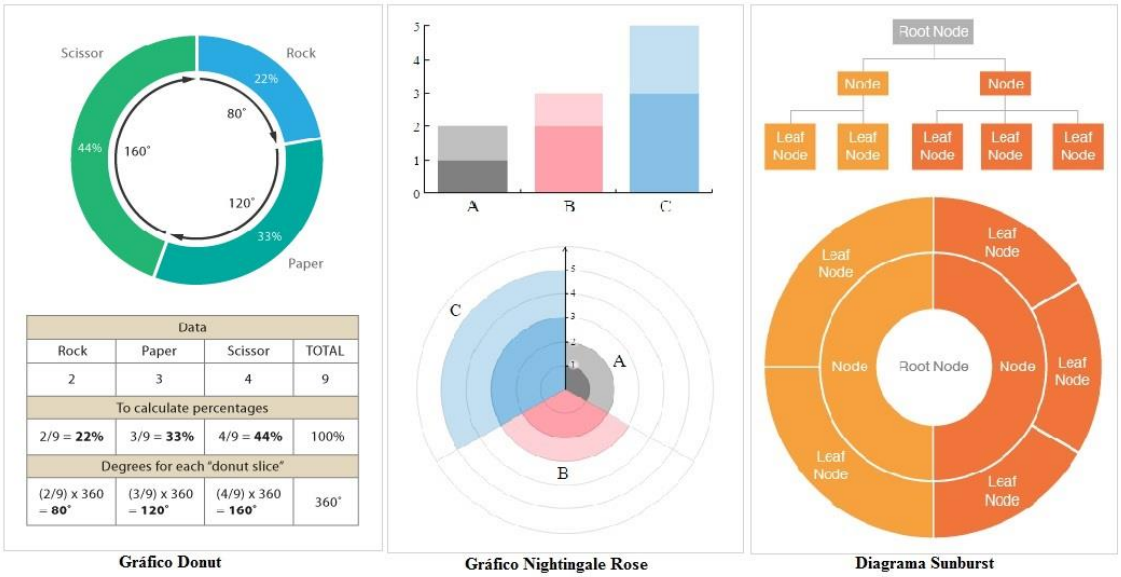


Gráfico Donut | Gráfico Nightingale Rose | Diagrama Sunburst

Fonte: Adaptado pelo autor

3.1.3 Gráfico de linha

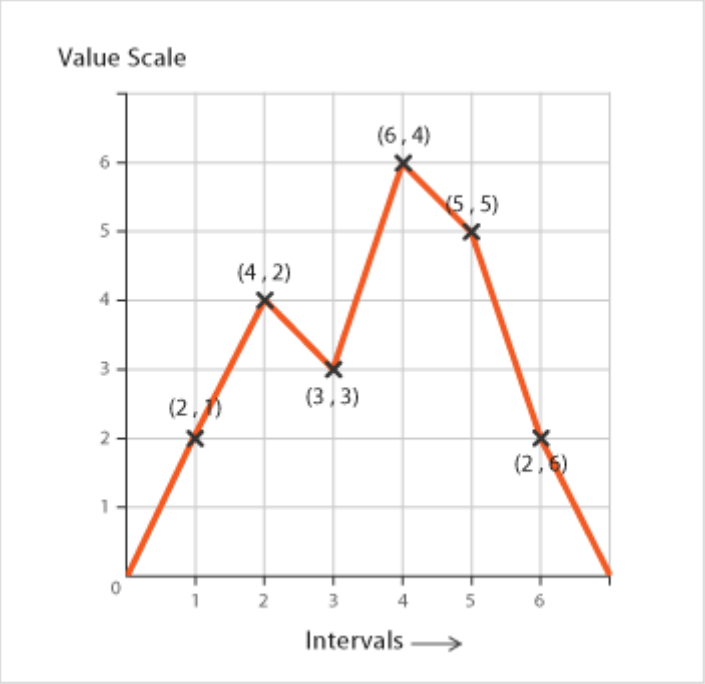
Gráficos de linha são usados para exibir valores quantitativos ao longo de um intervalo ou intervalo de tempo contínuo. É mais frequentemente usado para mostrar tendências e relações (quando agrupa várias linhas). (datavizcatalogue.com, 2016)

Gráficos de linha são construídos primeiro desenhando os pontos de dados (*data points*) em uma grade de coordenadas cartesianas, em seguida, conecta-se uma linha entre esses pontos. Normalmente, o eixo y tem um valor quantitativo, enquanto o eixo x tem uma categoria ou escala. Os valores negativos podem ser apresentados abaixo do eixo x. (datavizcatalogue.com, 2016)

SAS (2016) aponta um fator importante a ser levado em consideração. Não se deve escolher um gráfico de linhas para exibição de um conteúdo somente por que se têm *data points.* Em vez disso, deve-se observar o número de pontos de dados que se está trabalhando, pois este pode ditar o melhor visual para se usar. Por exemplo, se você tem 10 pontos de dados para exibir, a maneira mais fácil de entender esses 10 pontos pode ser simplesmente incluí-

los em uma ordem e mostrá-los usando uma tabela. A Figura 30 mostra um exemplo de um gráfico de linhas.

Figura 30 - Anatomia de um Gráfico de Linhas

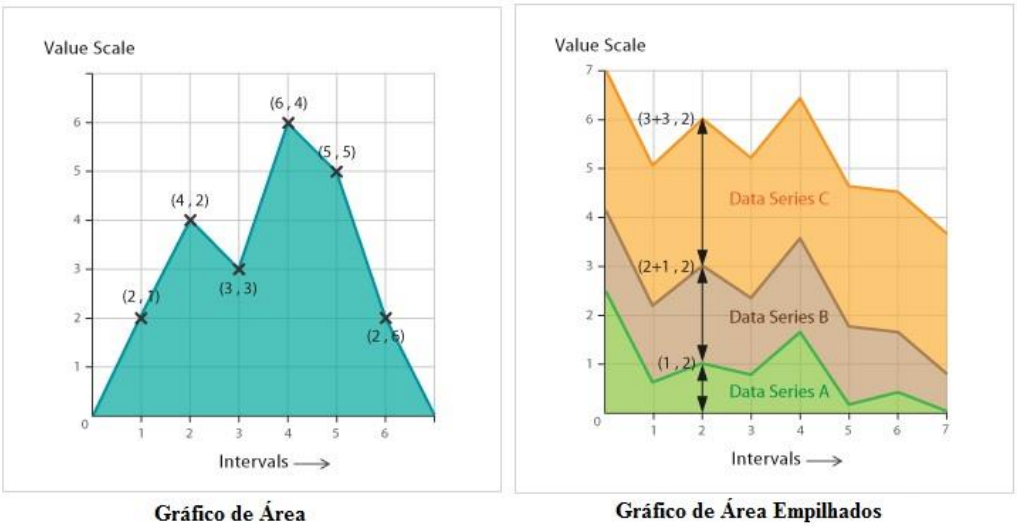


Fonte: http://www.datavizcatalogue.com/methods/line_graph.html

3.1.3.1 Gráficos Similares

Há variações do gráfico de linhas conforme aponta o portal datavizcatalogue.com (2016) que podem ser conferidos na Figura 31.

Figura 31 - Tipos de Gráficos de Linhas



Fonte: Adaptado pelo autor

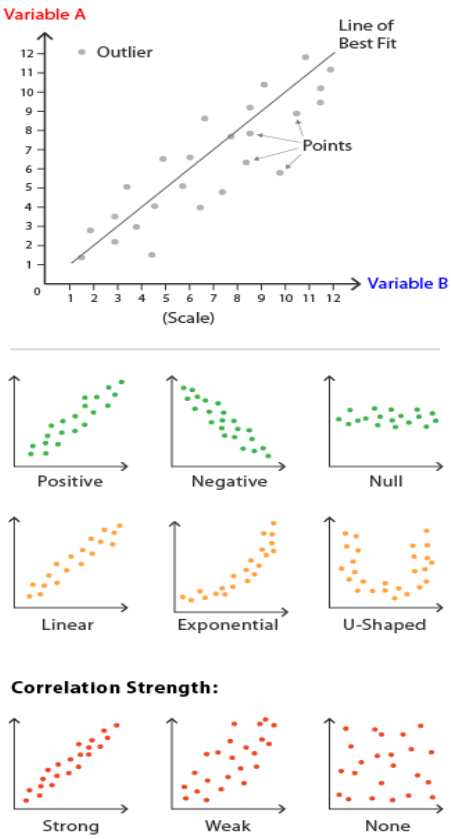
3.1.4 Gráficos de Dispersão

Também conhecido como *Scatter Plot*, Gráfico de Pontos, Gráfico X-Y, Gráfico de Dispersão ou Diagrama de Dispersão usa uma coleção de pontos distribuídos em coordenadas cartesianas para mostrar valores de duas variáveis. (datavizcatalogue.com, 2016)

SAS (2016) descreve este gráfico como sendo um gráfico bidimensional que mostra a junção de dois itens de dados, onde este ponto de junção indica o valor de cada observação. SAS ainda explica que *scatter plot* é muito utilizado para examinar correlação ou dependência entre variáveis. Por exemplo a relação entre lucro e receita que pode exibir uma correlação onde a medida que a receita aumenta o lucro também aumenta, caracterizando assim uma correlação positiva.

Datavizcatalogue.com (2016) mostra que há vários tipos de correlação como: positivos onde os valores aumentam em conjunto, negativo, ou seja, um valor diminui à medida que os outros aumentam nulo, que não há correlação linear exponencial e em forma de U. A força da correlação pode ser determinada pelo modo como os pontos são distribuídos, quanto mais próximos uns dos outros mais forte é a correlação. Pontos que acabam muito fora do *cluster* geral dos pontos são conhecidos como valores atípicos (*outliers*). A Figura 32 mostra exemplos de gráficos de dispersão com diferentes correlações.

Figura 32 - Anatomia de um Gráfico de Dispersão

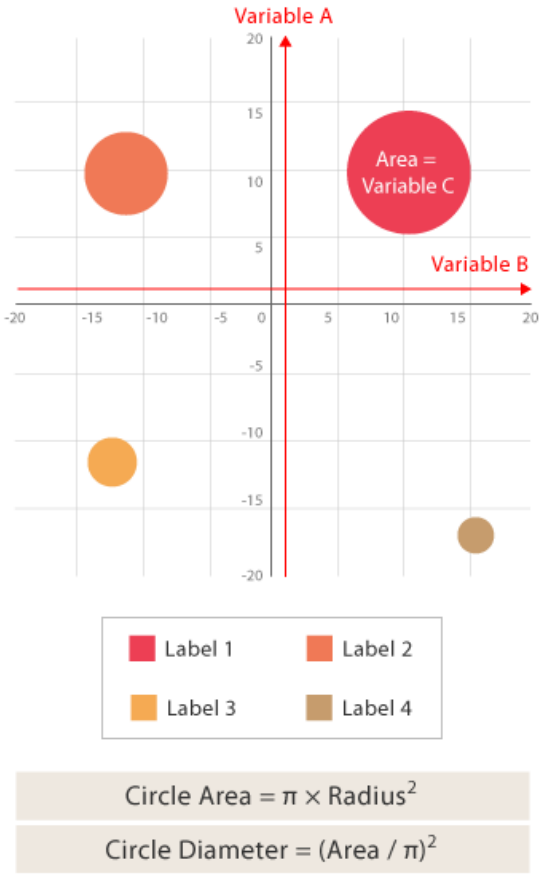


Fonte: http://www.datavizcatalogue.com/methods/scatterplot.html

3.1.4.1 Gráficos Similares

Há apenas uma variação do gráfico de dispersão conforme aponta o portal datavizcatalogue.com (2016) que pode ser conferida na Figura 33.

Figura 33 - Gráfico de Bolhas



Fonte: http://www.datavizcatalogue.com/methods/bubble_chart.html

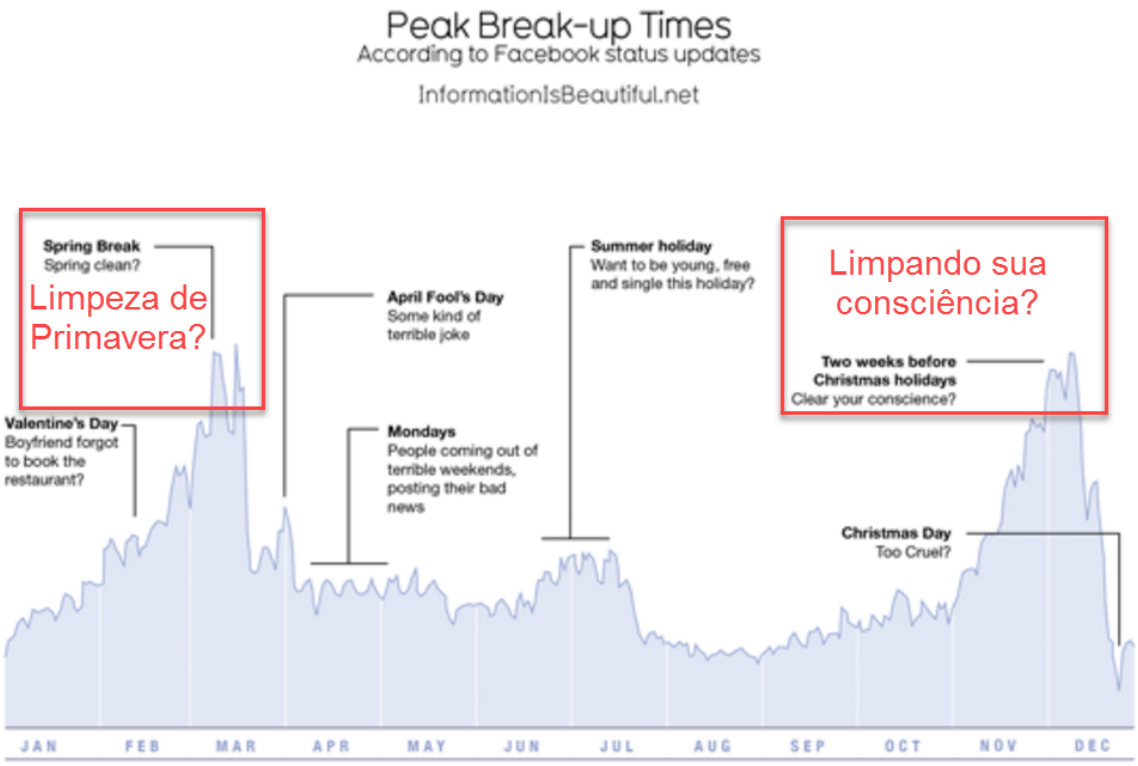
3.1.5 Gráficos Interativos

Quando se fala em interatividade de gráficos, pode-se dizer que há um toque de arte no que tange esta área da visualização de dados. David McCandless é um jornalista de dados britânico e designer de informação que em julho de 2010 apresentou uma palestra para o site TED Talk, intitulada *The beauty of data visualization*. Além da afirmação de que visualizar dados de forma gráfica ajuda a nos concentrar em partes específicas, identificar padrões, focar no que é importante e descobrir informações ocultas nos dados, visualizar gráficos pode ser algo muito legal e interessante. (McCandless, 2010)

Na Figura 34 pode-se ver que em algumas épocas do ano há um aumento no número de trocas de status de relacionamento no *Facebook*. Um exemplo é

a semanas do *spring break*[3], *o gráfico mostra o maior pico nesta época do ano que pode significar que muitos casais iniciam relacionamento, talvez com pessoas que conheceram nestas semanas de férias, enquanto outros seguem o caminho oposto, talvez para aproveitarem esse período*.

Figura 34 - Pico de rupturas/inicio de relacionamentos de acordo com mudança de status de relacionamento no facebook, 2008



Fonte: http://www.informationisbeautiful.net/2010/peak-break-up-times-on-facebook/

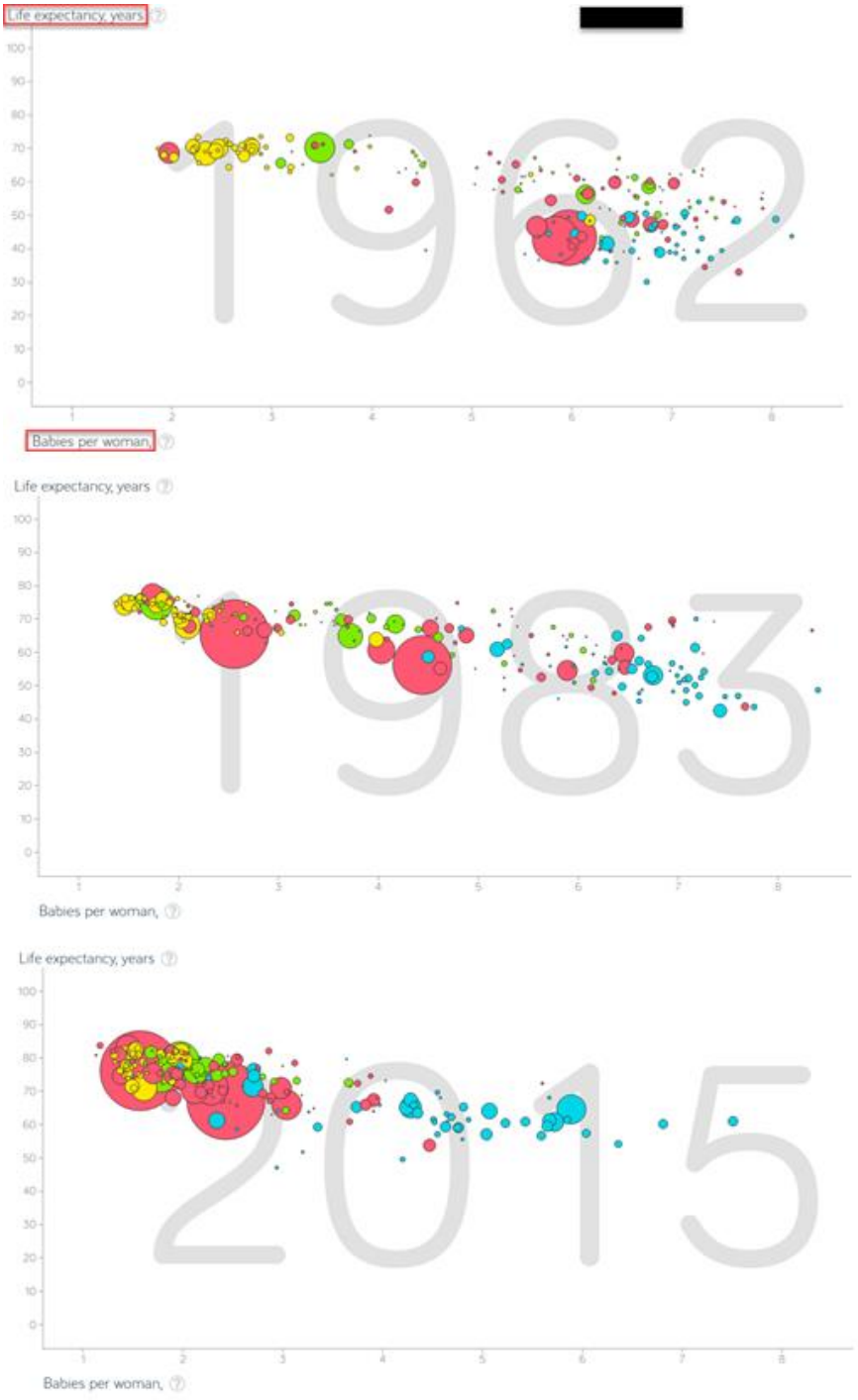
Porém, neste gráfico da Figura 34, não há muito campo para exploração de interatividade. Para esse conjunto de dados, mostrar o período de 12 meses, enfatizando os picos de mudança de status é uma forma eficiente de transmitir a informação desejada.

O professor Hans Rosling, especialista em saúde global e visionário de dados, mostrou também em uma palestra para o site TED Talk muitos recursos disponíveis em uma ferramenta desenvolvida por ele e outros colaboradores, o gapminder. Rosling demonstrou o quanto a interatividade em uma visualização, quando bem explorada, pode ser a melhor ferramenta para transmitir dados.

[3] O *Spring Break* é o período de uma semana em que as escolas e universidades americanas dão férias no início da primavera. (Misura,2014)

A sequência de imagens contidas na Figura 35, mostra a mudança ocorrida no mundo entre 1962 e 2015. O número de filhos diminui em muitos países ao mesmo tempo que a expectativa de vida aumentou.

Figura 35 - Expectativa de vida X Filhos por mulher



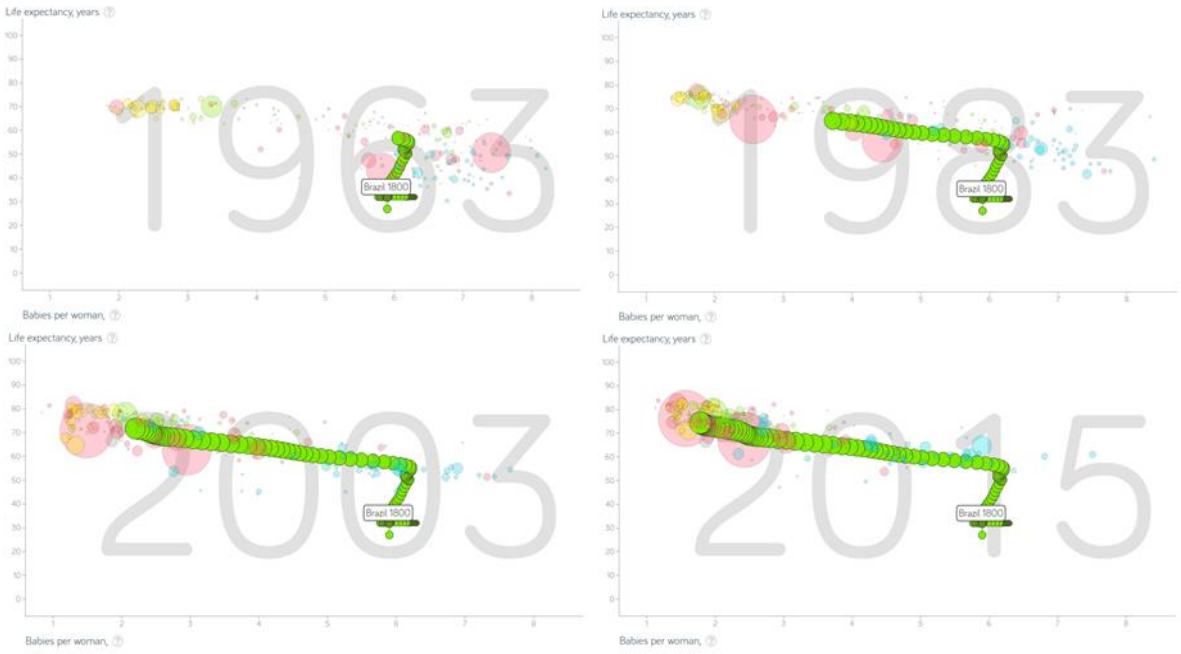
Fonte: Adaptado pelo autor

Essa mudança fica muito evidente, o padrão de aumento de expectativa de vida ao passo que números de filhos diminui fica ainda mais notável quando

visto interativamente na ferramenta do professor Rosling. Além de acompanhar essa evolução a nível mundial, podemos selecionar um país e comparar sua evolução, conforme mostra a Figura 36.

Nesta figura podemos ver que o Brasil, assim como os demais países do mundo, teve a expectativa aumentada em 1963 de 60 anos para algo perto de 80 anos em 2015. Enquanto o número de filhos por mulher diminui drasticamente, como visto na Figura 36, de 6 em 1963 para 1,5 em 2015.

Figura 36 - Expectativa de vida X Filhos por mulher, Brasil



Fonte: Adaptado pelo autor

3.2 Considerações finais

Como visto através de exemplos e descrito por alguns autores, visualizar os dados em forma de gráficos e imagens muitas vezes se torna mais fácil e pode ajudar o trabalho de identificar padrões até então desconhecidos em uma massa de dados ou instigar um pesquisador a seguir um determinado rumo na investigação desse mesmo conjunto de dados.

Porém, não há diretrizes ou regras que definem qual tipo de visualização é a correta para um dado conjunto de dados. Cabe ao analista entender o *dataset,* definir um grupo alvo para qual os dados serão apresentados e de que forma esse grupo consome os dados para então definir qual técnica será utilizada.

Neste trabalho não foram apresentadas todas as técnicas possíveis para se construir visualizações, pois como visto nos exemplos do item 3, é possível inventar novas maneiras de apresentar os dados, misturando imagens, gráficos e, se necessário outros elementos conforme a criatividade de quem for construir uma visualização. A seção 3.1.5 mostra como é importante entender os dados a serem exibidos e de que forma mostrá-los. Apesar de gráficos dinâmicos serem extremamente úteis, a simplicidade de um gráfico estático pode exibir de forma eficaz e simples toda a informação necessária para um determinado *dataset*. As técnicas aqui apresentadas servirão de base para se construir as visualizações necessárias para conclusão deste trabalho.

4. Ferramenta de Visualização

Para o desenvolvimento da ferramenta de visualizações utilizou-se a base de dados fornecida pela INCA[4]. Todos os arquivos de dados foram importados para dentro de uma variável *(data frame)* em R para então serem manipulados. Esse procedimento foi necessário para tornar possível uma melhor manipulação dos dados e então geração das visualizações necessárias, utilizando bibliotecas já disponíveis na linguagem, como é o exemplo do pacote Shiny[5] que é a base da interface gráfica[6].

4.1 Pacote Shiny

Conforme informado na introdução do trabalho, a linguagem R foi escolhida para ser a tecnologia principal utilizada na elaboração da ferramenta de visualização. Além da linguagem R, o pacote shiny é utilizado para construção da interface gráfica desta aplicação.

Em conjunto com o shiny, utilizou-se o pacote shinyDashboard. Este pacote fornece um tema montado sobre o Shiny, facilitando a criação de aplicações em estilo dashboard. (Chang *et. al.*, 2017). A união destes dois pacotes proporcionou a criação da aplicação contendo uma coluna lateral, à esquerda, para seleção dos parâmetros e a disposição dos gráficos dentro de *boxes* na tela principal, à direita, podendo-se assim ocultar certos *boxes*, se necessário.

A biblioteca Shiny possui muitas funções já disponíveis para manipulação e criação de elementos na interface de forma simplificada. Porém, uma das características que fizeram com que essa tecnologia fosse utilizada no trabalho é a solução *open source* conhecida como Shiny *server*[7]. Essa solução fornece uma plataforma na qual pode-se hospedar vários aplicativos Shiny em um único servidor, cada um com seu próprio URL ou porta. Assim os usuários da

[4] Página para download: https://irhc.inca.gov.br/RHCNet/visualizaTabNetExterno.action
[5] Página com todas informações sobre o pacote: http://shiny.rstudio.com/
[6] http://ceted.feevale.br:3838/inca Link para a aplicação.
[7] Página da plataforma: https://www.rstudio.com/products/shiny/shiny-server/

ferramenta podem acessa-la de forma online, da mesma maneira que acessariam um site qualquer através de um browser.

4.2 Base de dados

O INCA é o responsável pela base de dados utilizada neste trabalho. Os dados são primeiramente obtidos pelas unidades hospitalares que prestam assistência de alta complexidade em oncologia no Sistema Único de Saúde (SUS) e tem como principal fonte os prontuários médicos de pacientes com câncer atendidos nessas unidades. Após essa primeira captação de dados, toda a informação é reunida e consolidada em bases estaduais de responsabilidade das secretarias estaduais de saúde. (INCA,2017)

O papel do INCA é de reunir a informação fornecida pelas secretarias de saúde, retirar duplicidades para então gerar uma base nacional. Essa base contém informações referentes aos pacientes diagnosticados com a doença entre os anos de 1985 a 2015 (sem informações para o ano de 1987). A base está dividida por ano, sendo que cada paciente está inserido apenas uma vez na base no ano da sua primeira consulta no momento em que o câncer foi diagnosticado.

A base é disponibilizada em arquivos no formato dbf, sendo um arquivo para cada ano. Cada arquivo contém apenas o cabeçalho com o nome de cada atributo e os dados, similar a uma planilha de *Excel*. A Figura 37 mostra os 14 primeiros atributos de um dos arquivos sendo visualizados na ferramenta *Excel*.

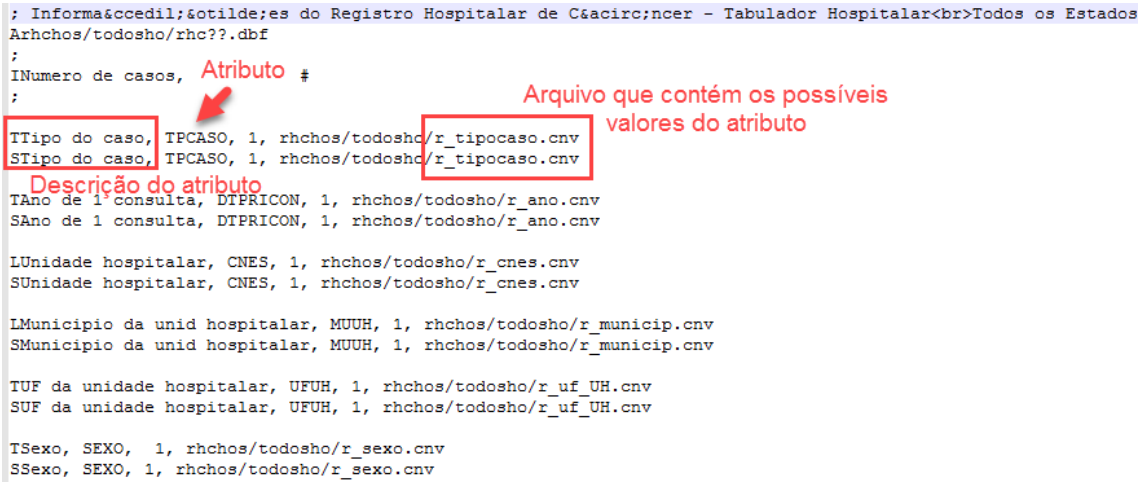
Figura 37: Exemplo de um arquivo dbf da base

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	TPCASO	SEXO	IDADE	LOCALNAS	RACACOR	INSTRUC	CLIATEN	CLITRAT	HISTFAMC	ALCOOLIS	TABAGISM	ESTADRES	PROCEDEN	ANOPRIDI
2	1	2	046	99	2	4	0	17	9	9	9	PE	2611309	03/08/2015
3	1	1	073	PE	4	2	0	17	9	9	9	PE	2611606	07/01/2015
4	1	2	076	PE	4	2	0	17	9	9	9	PE	2610707	24/03/2015
5	1	1	068	PE	4	2	0	17	9	9	9	PE	2607901	21/08/2015
6	1	2	083	PE	4	2	0	17	9	9	9	PE	2610707	09/02/2015
7	2	1	043	PE	4	6	0	88	9	9	9	PE	2611606	05/02/2015
8	1	2	041	PE	4	4	0	17	9	9	9	PE	2604106	04/02/2015
9	1	2	050	PE	4	1	0	17	9	9	9	PE	2613909	22/01/2015
10	1	2	034	PE	4	3	0	17	9	9	9	PE	2609600	10/03/2015
11	1	2	060	PE	4	2	0	17	9	9	9	PE	2604106	28/01/2015
12	1	1	060	PE	4	4	0	17	9	9	9	PE	2611606	19/02/2015
13	1	1	080	PE	4	2	0	17	9	9	9	PE	2613909	15/06/2015
14	1	1	064	PE	4	3	0	17	9	9	9	PE	2611606	06/05/2015
15	1	2	055	PB	4	2	0	17	9	9	9	PE	2511202	08/04/2015
16	1	2	065	PE	4	2	0	17	9	9	9	PE	2611606	15/10/2012
17	1	2	026	AL	4	3	0	17	9	9	9	PE	2702108	04/02/2015
18	1	1	033	PE	4	2	0	17	9	9	9	PE	2614204	04/02/2015
19	2	2	080	PE	4	3	0	17	9	9	9	PE	2611606	15/06/2007

Fonte: Adaptado pelo autor

Além dos arquivos dbf, o INCA disponibiliza junto com a base arquivos de apoio. Estes arquivos estão no formato cnv e descrevem os possíveis valores de cada atributo contido nos arquivos dbf. Há também um arquivo intitulado rhcGeral.def que indica o significado de cada atributo e qual arquivo de apoio que contém estes possíveis valores. Este arquivo funciona como um índice para que seja possível localizar o arquivo referente a cada atributo uma vez que os arquivos cnv têm nomes diferentes dos atributos. A Figura 38 ilustra a estrutura do arquivo rhcGeral.def.

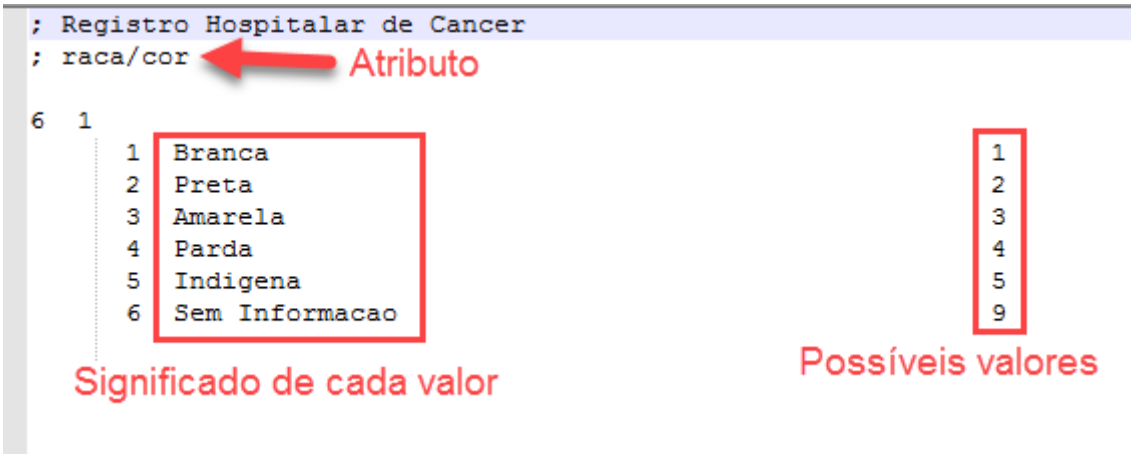
Figura 38: Estrutura arquivo rhcGeral.def



Fonte: Adaptado pelo autor

A Figura 39 demonstra a estrutura de um dos arquivos cnv, r_racacor.cnv, que descreve o atributo RACACOR da base de dados. Como visto na Figura 37, os arquivos dbf contém apenas números e datas por isso é necessário utilizar os arquivos cnv para identificar o significado de cada valor.

Figura 39: Estrutura do arquivo r_racacor.cnv



Fonte:Adaptado pelo autor

Na mesma página de download da base de dados, o INCA ainda disponibiliza as notas técnicas da base que apresentam informações relevantes no uso da mesma. De acordo com essas notas foi utilizada a Classificação Internacional de Doenças para Oncologia (CID-O) para representar a topografia e morfologia do tumor. A partir de 2005, a codificação dos tumores passou a ser feita utilizando-se a 3ª edição da CID-O (CID-O/3). Nos anos anteriores, era utilizada a 2ª edição da Classificação Internacional de Doenças para Oncologia (CID-O/2). O INCA não fez qualquer compatibilização de CID-O/2 para CID-O/3.

A Classificação Internacional de Doenças (CID) foi criada pela Organização Mundial de Saúde (OMS) e tem por objetivo fornecer códigos relativos a classificação de todas as doenças oficialmente reconhecidas. Para a área da oncologia, é utilizada uma versão especifica da CID a CID-O. Conforme destacam as notas técnicas da base é utilizada ainda a segunda edição para a consolidação dos dados, porém, as alterações sofridas entre a versão 2 e 3 são pequenas. O manual da Classificação Internacional de Doenças para Oncologia, lançado em 2000, explica que a segunda e terceira edição são iguais, sendo que a última edição foi revisada e teve algumas modificações no capitulo de linfomas e leucemias, onde foram introduzidos novos termos e códigos. A nomenclatura acrescida nesta 3ª edição inclui a classificação REAL (Revisão Europea-Americana de Linfomas) e também a classificação para leucemias, sistema FAB (Franco Americana-Britânica). (OMS, 2000)

Essa classificação descreve 4 dos 44 atributos contidos na base que são LOCTUDET (Localização Primaria), LOCTUPRI (Localização Primaria mais especifica), TIPOHIST (Tipo Histológico) e LOCTUPRO (Localização Primaria Provável). Esses 4 atributos obedecem a classificação internacional descrita na CID-O.

Essa classificação fornece um código para cada tipo de tumor juntamente com um código morfológico. Por exemplo, neoplasia maligna do pulmão (ex. carcinoma) é representada pelo código C34.9 e um segundo código que representa a sua morfologia incluindo um último dígito (separado dos demais por uma barra), que identifica o comportamento biológico, que nesse caso seria M-8010/3 (INCA, 2017).

Os atributos de TNM (TNM), Estadiamento (ESTADIAM) e Estadiamento Grupo (ESTADIAG) referem-se à avaliação da extensão da neoplasia maligna

antes do tratamento. Para o estadiamento dos tumores é utilizada a Classificação de Tumores Malignos (TNM) da União Internacional Contra o Câncer - UICC. Para os casos anteriores a 2005 é utilizada a 5ª Edição do TNM e para os casos a partir do ano 2005 é utilizada a 6ª edição do TNM.

Conforme explica a 6º edição presente no site do INCA as taxas de sobrevida de acordo com o estadiamento da doença são diferentes quando a doença está restrita ao órgão de origem ou se estende para outros órgãos. Estadiar um caso de neoplasia maligna é importante por motivos como verificar a taxa de crescimento da doença, extensão, relação entre o tipo de tumor e hospedeiro. (INCA, 2016)

Com o intuito de padronizar essa classificação surgiram alguns sistemas de estadiamento. O sistema utilizado nesta base de dados, único que será explicado neste trabalho, é o TNM preconizado pela UICC. Esse sistema é baseado nas características do tumor primário (T), as características dos linfonodos das cadeias de drenagem linfática do órgão em que o tumor se localiza (N) e a presença ou ausência de metástases à distância (M). (INCA, 2016)

Cada parâmetro geralmente recebe uma nota de acordo com o que representa, por exemplo T pode variar de 0 a 4, N de 0 a 3 e M de 0 a 1. Alguns dos parâmetros como T e N ainda podem conter subclassificações como a, b, c ou “X” que é utilizado quando uma categoria não pode ser devidamente classificada. (INCA, 2016)

Após cada categoria ser avaliada e agrupada, essa combinação resulta no estadiamento da doença que pode variar de I a IV e ainda pode ser subclassificada em A e B, para expressar o nível de evolução da doença.

Essa classificação é bastante útil e muito abrangente, porém, ainda existem outras classificações criadas para tumores específicos, que não dependem da combinação de TNM para definir o estadiamento da doença. Essa prática muitas vezes é utilizada por grupos que estudam tumores específicos, onde desenvolver uma própria classificação é mais eficaz. Um exemplo é o uso de letras maiúsculas (A, B, C, D) como ocorre no estadiamento de tumores de próstata, bexiga e intestino. A utilização de sistemas diferentes não os tornam incompatíveis, mas sim mais detalhados. Mesmo sendo o sistema da UICC o

padrão para essa base, para alguns tumores podem haver diferentes valores dos descritos acima. (INCA, 2016)

Além dos atributos já citados a base conta com outros 39 atributos, definidos da seguinte maneira (INCA,2017):

- Tipo do caso (TPCASO): registra se o caso é analítico ou não analítico;
- Sexo (SEXO): contém os dados referentes ao gênero do paciente contendo os valores de Masculino ou Feminino;
- Idade (IDADE): indica a idade do paciente ao ser diagnosticada a doença;
- Local de Nascimento (LOCALNAS) que indica o estado de nascimento do paciente quando brasileiro já para estrangeiros no lugar da sigla do estado é colocado “EX”;
- Raça e cor (RACACOR): contém as seguintes categorias: branca, preta, amarela, parda, indígena e sem informação;
- Grau de Instrução (INSTRUC): refere-se ao nível de instrução do paciente onde as categorias contidas na base são: analfabeto, 1 º grau incompleto, 1º grau, 2 º grau, superior e sem informação;
- Clínica de 1º atendimento (CLIATEN): variável se refere ao serviço médico especializado responsável pela matrícula e atendimento inicial ao paciente no hospital;
- Clínica de Tratamento (CLITRAT): esta variável possibilita a identificação da clínica onde efetivamente foi iniciado o tratamento antineoplásico do paciente, se o tratamento foi realizado por mais de uma clínica considera-se a clínica que assumiu o papel primordial no tratamento;
- História Familiar de Câncer (HISTFAMC): atributo que mostra histórico de câncer na família exclusivamente aos parentes consanguíneos, ascendentes ou colaterais até segunda geração, ou seja, pais, irmãos, avós e tios;
- Alcoolismo (ALCOOLIS): essa variável se refere à história de consumo de bebida alcoólica, não apenas a situação atual, mas a ocorrência preponderante, sendo os valores como: sim, não, não se aplica e sem informação;

- Tabagismo (TABAGISM): de forma semelhante ao atributo de Alcoolismo essa variável avalia o uso de tabaco pelo paciente as categorias disponíveis são: sim; não; não se aplica e sem informação;
- Estado de Residência (ESTADRES): apresenta o estado de residência do paciente;
- Procedência (PROCEDEN): atributo que mostra o município onde o paciente vive;
- Ocupação (OCUPACAO): registra a ocupação do paciente;
- Ano do Diagnostico (ANOPRIDI): refere-se ao ano em que foi realizada a confirmação do diagnóstico de câncer do paciente;
- Origem do encaminhamento (ORIENC): esta variável se refere à origem do encaminhamento do paciente à Unidade Hospitalar. As categorias disponíveis são: SUS; não SUS; veio por conta própria e sem informação;
- Diagnóstico e Tratamento Anterior (DIAGANT): esta variável se refere ao estabelecimento do diagnóstico e das medidas terapêuticas específicas para o tumor, realizadas antes do paciente dar entrada no hospital.
- Base Diagnostico (BASMAIMP): refere-se ao exame sobre o qual foi estabelecido, com maior grau de certeza, o diagnóstico de câncer do paciente. As categorias disponíveis são: exame clínico e patologia clínica; exame por imagem; endoscopia; cirurgia exploradora/necropsia; citologia ou hematologia; histologia da metástase; histologia do tumor primário; sem informação;
- Exames para Diagnóstico (EXDIAG): variável que se refere aos exames relevantes para o diagnóstico e planejamento da terapêutica do tumor;
- Lateralidade (LATERALI): refere-se à lateralidade do tumor e somente é preenchida para tumores de órgão par, com objetivo de estudar a frequência de tumores em órgãos múltiplos, por exemplo: mama, pulmão, rim, entre outros. As categorias disponíveis são: direita; esquerda; bilateral; não se aplica e sem informação;
- Tumor Primário Múltiplo (MAISUMTU): atributo que mostra à ocorrência de mais de um tumor primário em um determinado órgão ou em órgãos diferentes. Os tumores podem ocorrer, simultaneamente ou não, em

diferentes localizações de um mesmo órgão ou em diferentes órgãos, com a mesma histologia ou não;

- Ano do Primeiro Tratamento (DTINITRT): variável que armazena o ano em que o primeiro tratamento foi aplicado;
- Primeiro tratamento recebido no hospital (PRITRATH): esse atributo armazena informação do primeiro tratamento recebido em hospital onde as possíveis categorias são nenhum; cirurgia; quimioterapia (QT); radioterapia (RXT); hormonioterapia (HT); transplante de medula óssea (TMO); categorias que se referem à combinação dessas modalidades de tratamentos; outros procedimentos terapêuticos; sem informação;
- Razão para não tratar (RZNTR): armazena o motivo para não tratar a doença com as seguintes categorias: recusa do tratamento; doenças avançadas; falta de condições clínicas; outras doenças associadas; abandono do tratamento; complicações do tratamento; óbito; outras; não se aplica e sem informação;
- Estado da doença ao final do primeiro tratamento (ESTDFIMT): mostra o estado da doença ao final do primeiro tratamento proposto, onde as categorias disponíveis são: sem evidência da doença (remissão completa); remissão parcial; doença estável; doença em progressão; fora de possibilidade terapêutica; óbito; não se aplica e sem informação;
- Estado conjugal (ESTCONJ): esta variável se refere ao estado conjugal atual do paciente, e não deve ser confundida com estado civil, ou seja, não é a situação legal do casal. As categorias disponíveis são: casado, solteiro, desquitado/separado/divorciado, viúvo e sem informação;
- Ano da Triagem (ANTRI): atributo referente ao ano do primeiro contato do paciente na Unidade Hospitalar no processo relacionado ao tumor;
- Ano da Primeira consulta (DTPRICON): variável para armazenar ano da primeira consulta realizada referente ao tumor;
- Unidade Hospitalar (CNES): atributo para guardar código da unidade hospitalar que prestou assistência oncológica e cadastrou o paciente;
- Município da unidade hospitalar (MUUH): variável que armazena o município da unidade hospitalar que prestou assistência oncológica;

- UF da unidade hospitalar (UFUH): responsável por armazenar o estado da unidade hospitalar;
- Data Óbito (DATAOBITO): armazena a data do óbito do paciente;
- Data do Diagnostico (DTDIAGNO): data em que foi realizada a confirmação do diagnóstico de câncer do paciente;

A três ultimas variáveis não apresentam informações sobre seu conteúdo que são DTTRIAGE, DATAPRICON, DATAINITRT.

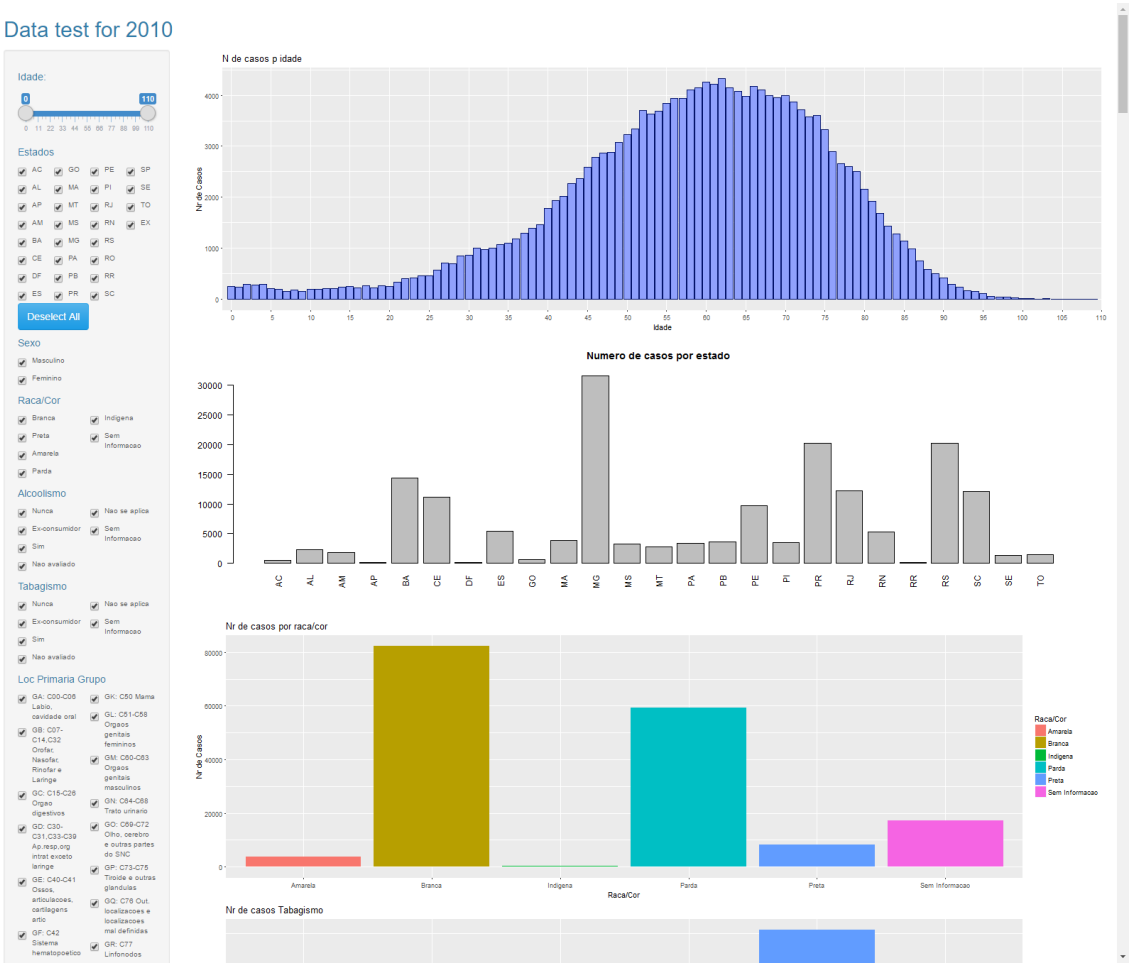
Foram selecionados os atributos IDADE, UFUH, SEXO, RACACOR, LOCTUDET, TABAGISMO, ALCOOLIS, ESTADIAG, DTPRICON para disponibilização como filtros dos dados na ferramenta. Esses atributos foram recomendados por pesquisadores da área da saúde que conhecem a base de dados.

4.3 Construção da Ferramenta para Visualização da Base

Para construção da ferramenta utilizou-se a linguagem de programação R pela sua capacidade de gerar visualizações das mais variadas formas e com muitas bibliotecas já disponíveis, tudo de forma gratuita e *open source*. Em um primeiro modelo da ferramenta combinou-se a biblioteca shiny, utilizada para criar a estrutura do aplicativo e a biblioteca ggplot2, utilizada para gerar os gráficos, além de algumas funções nativas da linguagem R usadas também para criação de gráficos.

No primeiro modelo foram geradas algumas visualizações utilizando apenas um arquivo da base do ano de 2010(150 MB, 233.814 registros) com a possibilidade de seleção dos filtros dos dados utilizando idade, estados, sexo, utilização de tabaco ou álcool, raça dos pacientes e local primário do tumor por grupos. A Figura 40 mostra um pedaço do primeiro protótipo.

Figura 40: Tela do Primeiro Protótipo da Ferramenta com informações da base para o ano de 2010



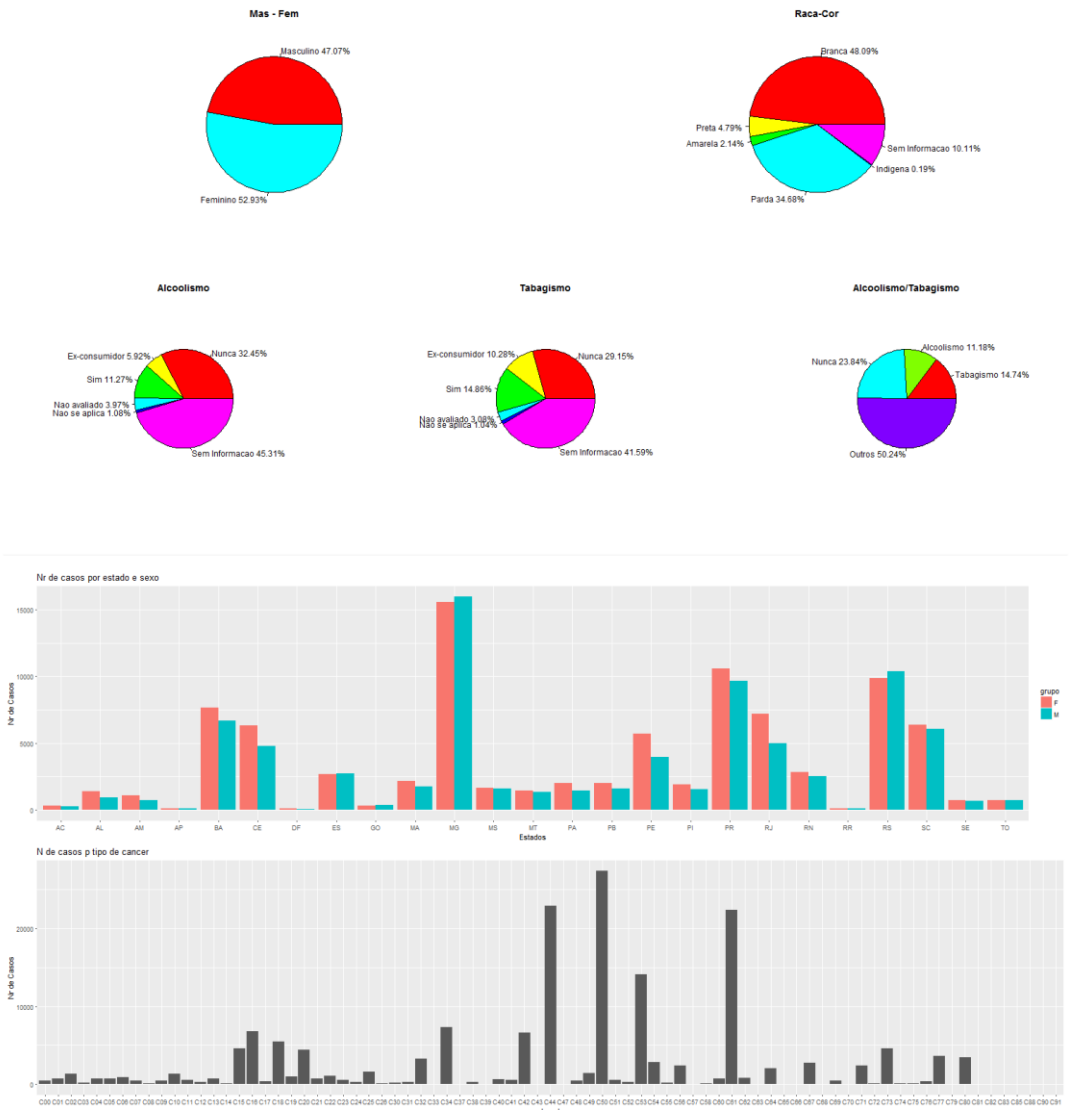
Fonte: Adaptado pelo autor

Este primeiro protótipo mostrou que havia muito potencial a ser explorado utilizando-se a linguagem R e suas bibliotecas. Porém, ao mesmo tempo encontrou-se alguns problemas de utilização da ferramenta. Neste primeiro modelo, ao selecionar quais informações seriam exibidas ou excluídas utilizando-se os *checkboxes* à esquerda da tela, o modelo recarregava todos os gráficos para cada alteração feita, ou seja, se o usuário desmarcasse as opções masculino, branca, parda a ferramenta carregaria os gráficos três vezes, mesmo o usuário tendo desmarcado todas opções praticamente ao mesmo tempo. Isso ocasionava em um tempo de renderização da informação na tela muito grande, tornando a experiência de utilização nada agradável.

Outros detalhes menores como aparência de alguns gráficos, formatação da informação exibida, como foi o caso dos gráficos de pizza, não ficaram bons. A Figura 41 mostra como a aparência de um modo geral da ferramenta ainda

não estava agradável. Os gráficos continham aparências muito distintas, os gráficos de pizza apresentavam algumas legendas sobrepostas e as cores não eram agradáveis aos usuários. Assim como o ultimo gráfico da Figura 41, os tons de cinza não tornam a imagem instigante, a quantidade de informação e a distância de algumas barras das legendas laterais tornavam difícil identificar o valor correspondente a cada barra.

Figura 41: Exibição da Informação de um Modo não Satisfatório



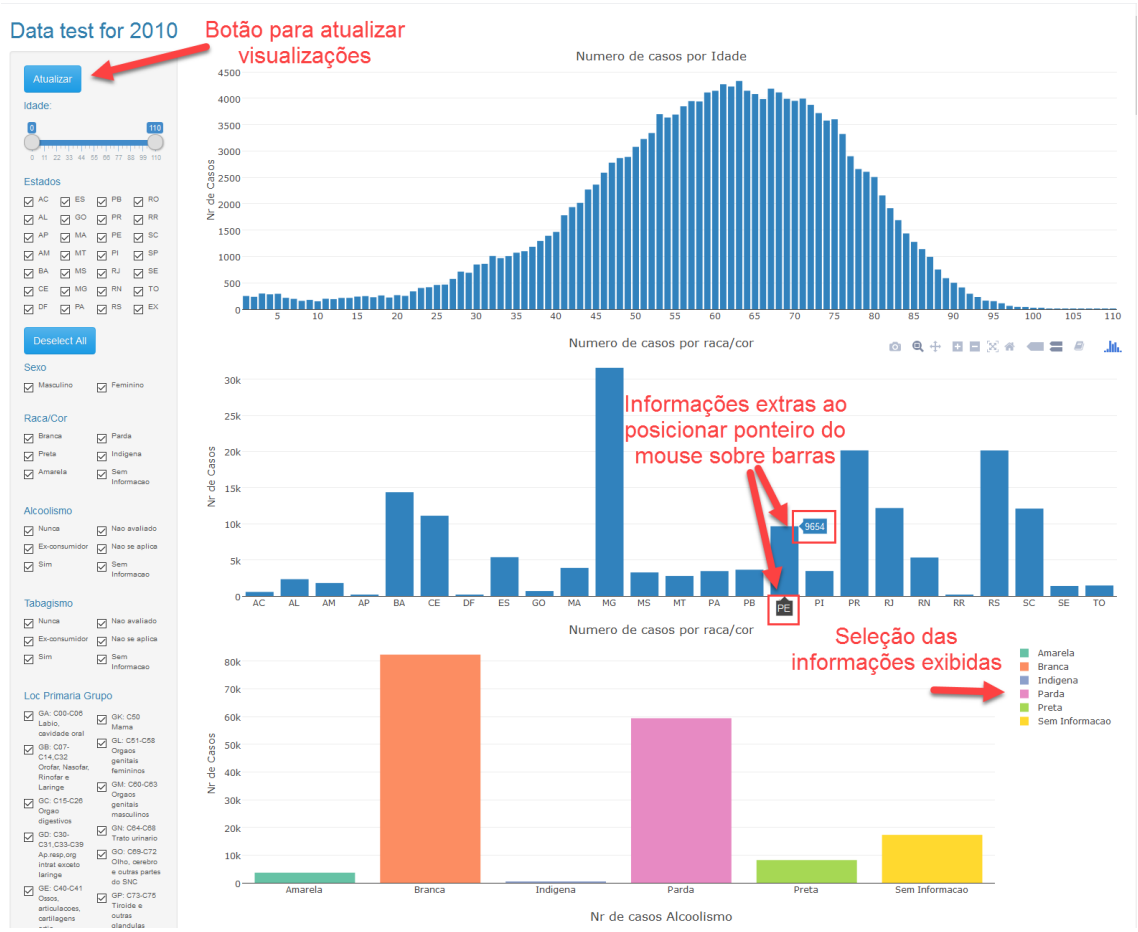
Fonte: Adaptado pelo autor

Em busca de resolver estas questões e tornar a experiência de uso da ferramenta mais agradável buscou-se uma solução para questão da renderização da informação. Após pesquisas sobre possíveis métodos para correção deste problema, optou-se por implementar um botão de atualização dos

gráficos. Após o usuário fazer todas as alterações necessárias nos parâmetros de exibição da informação, o mesmo deverá clicar neste botão para então ter todos os gráficos gerados novamente respeitando os parâmetros selecionados.

Essa modificação permite que o usuário selecione todos os parâmetros desejados e faça a renderização dos gráficos somente após terminar as modificações, isso evita com que a aplicação atualize os gráficos múltiplas vezes de forma desnecessária. A Figura 42 mostra a implantação deste botão em uma versão mais recente da ferramenta.

Figura 42: Nova Versão da Ferramenta já com Botão para Atualizar as Visualizações



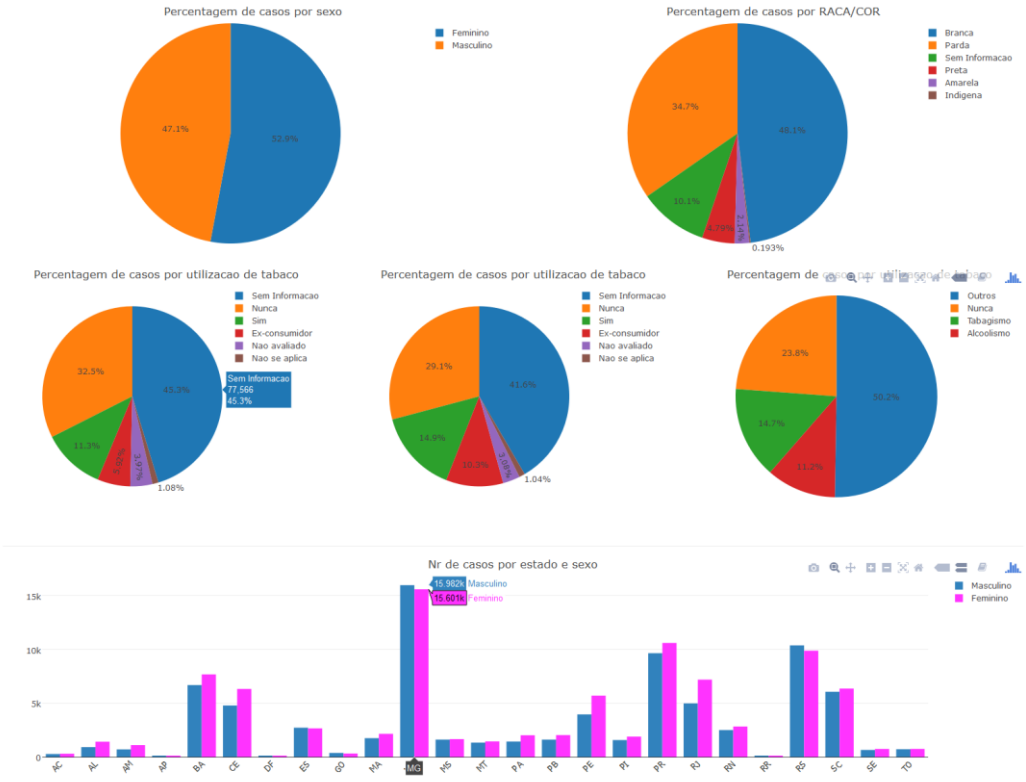
Fonte: Adaptado pelo Autor

Buscando resolver problemas de exibição de alguns gráficos e melhorar a aparência da aplicação, encontrou-se a biblioteca Plotly. Essa biblioteca trouxe a possibilidade de disponibilizar novas funcionalidades para o usuário da ferramenta. Conforme mostra a Figura 42, além de gráficos mais padronizados, foi possível implementar funções para mostrar mais informações sobre cada barra de um gráfico. O segundo gráfico apresenta o estado e o número de casos de câncer ao se posicionar o ponteiro sobre a barra.

Outra funcionalidade obtida com essa nova biblioteca é a possibilidade de remover barras especificas de um gráfico, aproximar ou movimentar o gráfico dentro de sua área de exibição.

A biblioteca Plotly trouxe novas funções para a ferramenta e ainda melhorou de forma expressiva a aparência da aplicação. Todos os gráficos foram gerados utilizando a nova biblioteca, o que agregou uma melhor padronização na construção e exibição dos gráficos.

Figura 43: Nova Versão da Ferramenta Utilizando Plotly



Fonte: Adaptado pelo Autor

A melhora foi muito expressiva nos gráficos de pizza, assim como nos gráficos de barra. Os gráficos de pizza também ganharam a funcionalidade de excluir fatias especificas das visualizações, além da melhor aparência na forma geral da exibição de cada gráfico e na informação ali contida.

A Figura 43 exibe estas melhoras, assim como a funcionalidade de poder ter informações do número de casos na comparação entre o número de pacientes com a doença divididas por estado e ainda por sexo, conforme apresentado no ultimo gráfico da Figura 43.

Até a versão da ferramenta apresentada na Figura 43, não haviam sido feitas muitas mudanças no código fonte e na forma visual da aplicação. O acesso

a base de dados ainda acontecia através da leitura dos arquivos baixados no site do INCA, porém, na versão final da ferramenta foram feitas modificações expressivas no código fonte, na estrutura visual e também na maneira de armazenar e acessar as informações da base.

A ferramenta prevê a necessidade de acessar os dados sobre câncer de todos os anos disponíveis na base do INCA. Desta forma, foi transferido o conteúdo de cada arquivo de forma separada para dentro da aplicação em cada execução e gerada as visualizações (1.57 GB). O procedimento de renderização dos gráficos foi lento. Afim de agilizar esse processo e também melhor manejar os dados contidos em cada arquivo, inseriu-se todo o conteúdo dos 30 arquivos em um banco de dados PostgreSql.

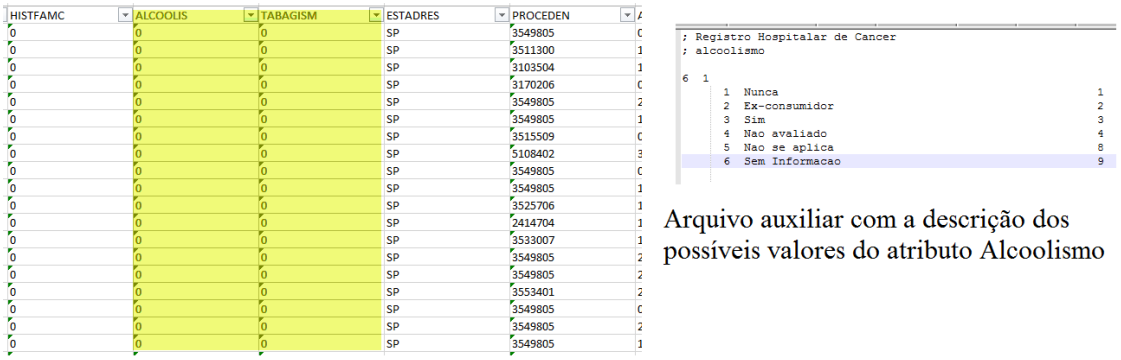
Utilizou-se o script contido no Apêndice B para importar o conteúdo de todos os arquivos para dentro da base. O script foi construído para migrar um arquivo por vez, assim o nome do arquivo que seria migrado para a base foi inserido no script e o mesmo foi executado 30 vezes para completar toda a migração.

A nova tabela criada dentro do banco de dados, tb_inca, é praticamente um espelho dos arquivos do INCA a única diferença foi a inclusão de um novo atributo chamado id que foi criado para atrelar um identificador único para cada registro. Essa adição aos atributos da tabela foi feita pensando em futuras manutenções dos registros desta tabela e também possíveis relações que venham ser necessárias em caso de criação de novas tabelas.

Esse processo auxiliou na limpeza da base que foi necessária devido a existência de muitos dados inconsistentes com o que estava descrito nos arquivos cnv auxiliares. Essa inconsistência ficou evidente em uma das validações feitas na ferramenta. Ao verificar o número de registros por estado, notou-se que não haviam dados para o estado de São Paulo. Após investigação verificou-se que em todos os arquivos os registros de pacientes de São Paulo continham valores zero para os atributos de ALCOOLIS e TABAGISM e de acordo com os arquivos auxiliares para estes dois atributos os valores possíveis estão entre 1 a 4, 8 e 9. Como o zero não estava mapeado todos os valores foram substituídos por 9 que significam “Sem Informação”.

Com todo o conteúdo dentro de um banco de dados essa tarefa se tornou mais fácil, em função do uso de SQL. O Apêndice C mostra as *queries* realizadas para auditoria de outros atributos utilizados no trabalho.

Figura 44: Exemplo de Dados que Precisaram ser Manipulados



HISTFAMC	ALCOOLIS	TABAGISM	ESTADRES	PROCEDEN
0	0	0	SP	3549805
0	0	0	SP	3511300
0	0	0	SP	3103504
0	0	0	SP	3170206
0	0	0	SP	3549805
0	0	0	SP	3549805
0	0	0	SP	3515509
0	0	0	SP	5108402
0	0	0	SP	3549805
0	0	0	SP	3549805
0	0	0	SP	3525706
0	0	0	SP	2414704
0	0	0	SP	3533007
0	0	0	SP	3549805
0	0	0	SP	3549805
0	0	0	SP	3553401
0	0	0	SP	3549805
0	0	0	SP	3549805
0	0	0	SP	3549805

Exemplo de alguns registros para o estado de São Paulo

```
; Registro Hospitalar de Cancer
; alcoolismo

6  1
     1  Nunca                    1
     2  Ex-consumidor            2
     3  Sim                      3
     4  Nao avaliado             4
     5  Nao se aplica            8
     6  Sem Informacao           9
```

Arquivo auxiliar com a descrição dos possíveis valores do atributo Alcoolismo

Fonte: Adaptado pelo Autor

A Figura 44 ilustra o exemplo de alguns registros contendo o valor 0 (zero) para pacientes de São Paulo e o respectivo arquivo cnv que contém o significado de cada valor do atributo Alcoolismo. Essa mesma comparação foi realizada com os outros atributos, porém não foi tomado nenhuma ação corretiva, pois diferente dos registros de São Paulo, os demais problemas encontrados não representavam uma porcentagem alta em comparação ao número total de registros da base. Os erros encontrados para cada atributo e número de elementos problemáticos podem ser conferidos também no Apêndice C.

Outra ação realizada dentro do banco de dados foi a criação de uma nova tabela para acomodar os dados necessários para executar a tarefa de regressão linear. Essa tarefa necessita que os dados estejam organizados de maneira especifica e para evitar que essa organização fosse realizada através da aplicação a cada execução da ferramenta, todos os dados foram preparados e inseridos em uma nova tabela chamada tb_regressao. Isso facilita a recuperação da informação e agiliza o processo de inicialização da ferramenta.

A tabela criada para gerar as regressões contém 6 atributos que são:

Ano: armazena o ano referente aos registros coletados;

Grupo: mesmo sistema de agrupamento utilizado para filtrar os dados na seção dos gráficos dinâmicos. Este atributo registra o grupo de órgão em qual o câncer

se desenvolveu tendo os grupos separados de acordo com o arquivo auxiliar r_cido2t_agrupado.cnv;

Total_casos: coluna que armazena o número total de casos para determinado grupo;

Total_alcool: registro do número total de pacientes que consomem álcool e desenvolveram câncer em um dos órgãos do grupo descrito;

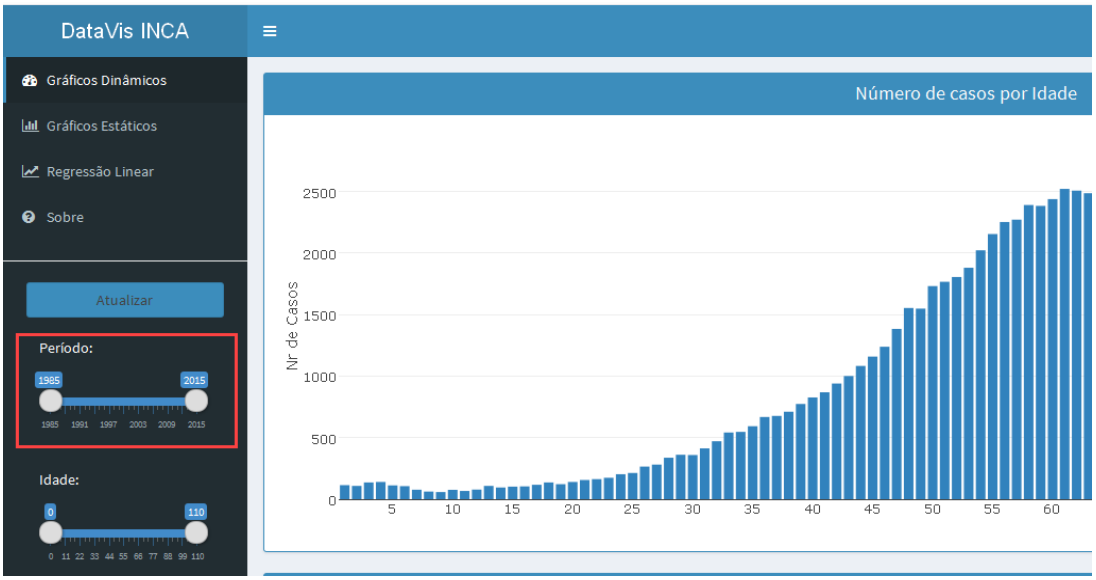
Total_tabagismo: registro do número total de pacientes que consomem tabaco e desenvolveram câncer em um dos órgãos do grupo descrito;

Total_alctab: registro do número total de pacientes que consomem álcool e tabaco, e desenvolveram câncer em um dos órgãos do grupo descrito;

Os atributos mencionados acima foram selecionados devido à grande discussão que há em torno do consumo de álcool e tabaco estarem diretamente ligados ao desenvolvimento da doença. Oferecer uma opção de gerar regressões lineares sobre estes dados oferece mais valor à aplicação.

O armazenamento da informação em um banco de dados gerou algumas alterações no código fonte da aplicação para acessar esse conteúdo. Junto foram implementadas novas funções como uma barra para seleção do período de amostragem dos dados. O usuário agora pode escolher o intervalo de tempo para exibir a informação. A possibilidade de selecionar os dados por estado ou região e o estadiamento do tumor por grupo também foram acrescidas. A Figura 45 mostra a nova barra para seleção do período de captura dos dados.

Figura 45: Nova Barra para Seleção do Período dos Dados



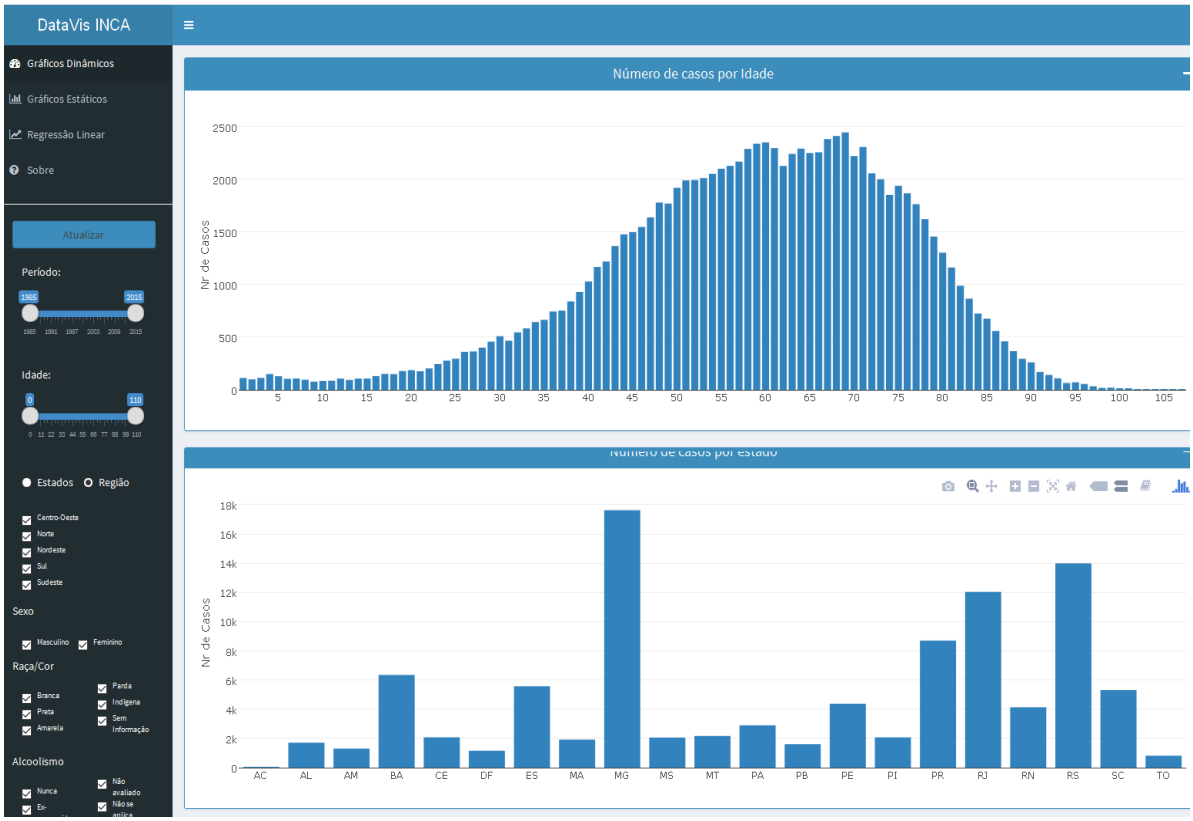
Fonte: Adaptado pelo Autor

A maior mudança foi na parte visual da aplicação. Com o uso da biblioteca *Shiny Dashboard*, a aparência da ferramenta ficou muito mais agradável e funcional. Abas foram adicionadas na barra lateral esquerda da tela para dividir os tipos dos gráficos. Ao selecionar os gráficos estáticos, o controle de seleção dos dados é ocultado e a tela substitui os gráficos da seção antiga pela nova.

As divisões das abas foram feitas da seguinte maneira:

- Gráficos Dinâmicos: Foram colocados nessa seção todos os gráficos que tem sua visualização alterada de acordo com a seleção dos parâmetros feitas pelo usuário (no menu à esquerda);
- Gráficos Estáticos: Essa seção comporta os gráficos que se mantém estáticos indiferente das seleções feitas pelo usuário, esses gráficos mostram informações da base de dados que comportam todo o período disponível;
- Regressão Linear: Seção dedicada ao uso de Regressão Linear, onde o usuário pode selecionar os parâmetros de tabagismo e alcoolismo para gerar a regressão;
- Sobre: Aba que contém informações do trabalho e da base de dados;

Figura 46: Versão Final da Ferramenta de Visualização

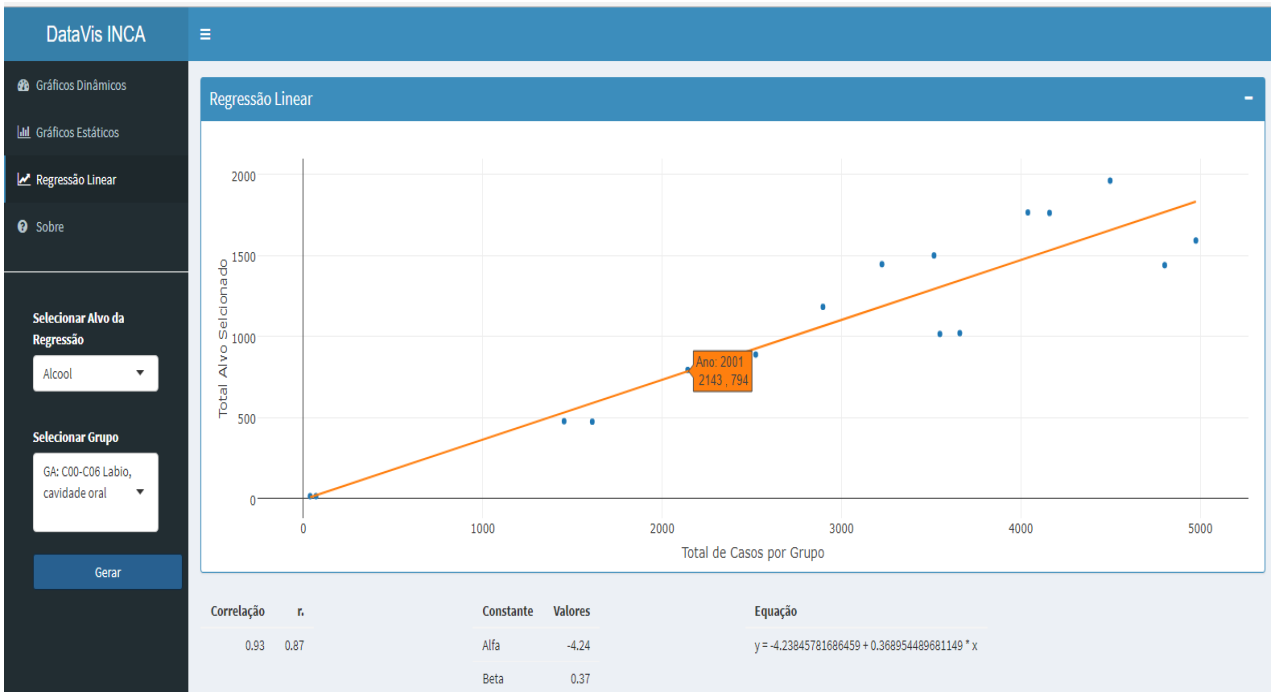


Fonte: Adaptado pelo Autor

A Figura 45 mostra a versão final da ferramenta. O uso da biblioteca Shiny Dashboard trouxe uma aparência mais profissional e uma melhor exibição da informação, permitindo o uso de abas para melhorar a organização de toda a aplicação.

Nesta última versão foi incluso a aba de Regressão Linear para comportar o gráfico referente a tarefa de mineração de dados de mesmo nome. A Figura 47 ilustra o gráfico gerado a partir da seleção dos parâmetros do grupo de órgãos que se deseja analisar e a variável alvo que pode variar entre Álcool, Tabagismo ou Ambos.

Figura 47: Aba de Regressão Linear



Fonte: Adaptado pelo Autor

O gráfico da figura 47 mostra a regressão gerada para o grupo de cânceres formados na cavidade oral buscando verificar a correlação para pacientes usuários de álcool. O valor da correlação para estes parâmetros é de 0.93, o que mostra uma forte correlação entre essas variáveis. Além do valor de correlação os usuários podem consultar os valores das constantes Alfa e Beta, assim como a equação da reta em questão. Por motivos de melhor aproveitamento do espaço de visualização, os valores de Alfa e Beta foram apresentados apenas com duas casas decimais mas para a equação utilizou-se o valor completo de cada constante para uma obter-se uma precisarão maior no resultado.

Na lateral da figura, pode-se ver os parâmetros selecionados para gerar a regressão. A Variável alvo selecionada foi Alcool e o grupo GA, que representa todos os cânceres desenvolvidos na cavidade bucal. No centro da imagem, ao posicionar o cursor do mouse sobre um ponto apresenta-se o ano referente aos dados daquela observação (2001), o número de pacientes usuários de álcool, (794) e o número total de casos registrados para o grupo no ano informado (2143).

Nesta versão foi implementado apenas a tarefa de regressão linear. Apesar do trabalho apresentar muitas técnicas que podem ser utilizadas sobre

esta base, ao decorrer do desenvolvimento do trabalho evidenciou-se que os pesquisadores que já conheciam a base do INCA não tinham conhecimento de toda a informação ali disponível. Viu-se então que havia uma necessidade maior de gerar visualizações e gráficos sobre os dados ali contidos do que propriamente trabalhar mais as técnicas de mineração. Dado ao restrito espaço de tempo para desenvolvimento do trabalho optou-se por disponibilizar mais opções de manipulação dos dados existentes na base.

A versão final da ferramenta foi apresentada para um pesquisador e profissional da área da saúde juntamente com um questionário para avaliação e coleta de *feedback* sobre a aplicação.

De acordo com a informação coletada com o profissional, a ferramenta é de fácil utilização e pode ser utilizada por pesquisadores para analisar os dados da base para visualizar padrões menos óbvios e delinear projetos e ações nessa área de pesquisa.

Referente a performance da ferramenta, o entrevistado não apontou quaisquer problemas para utilização. Mesmo havendo a necessidade de espera de alguns segundos para a primeira geração dos gráficos, a utilização da aplicação de um modo geral ainda é mais vantajosa que a análise dos mesmos dados utilizando ferramentas tradicionais disponíveis a usuários não ligados a área de computação que é o exemplo do Microsoft Excel.

Não foram apontadas desvantagens na utilização da aplicação. Por se tratar de algo inovador para a área, há muitas vantagens quando comparado as ferramentas existentes. Porém, foram apontadas muitas melhorias possíveis para ferramenta como a possibilidade de movimentar o menu lateral no sentido vertical, para que alterações nos parâmetros de seleção dos dados possam ser realizadas sem perder a visibilidade de algum gráfico especifico ou inclusão de mais alguns atributos da base e adição da possibilidade de gerar outras regressões.

5. CONCLUSÃO

O estudo bibliográfico realizado mostrou que o câncer apesar de ser uma doença antiga e extremamente agressiva ainda não possui uma cura que seja totalmente eficaz para todos os casos e indivíduos. Isso faz com que o estudo e pesquisa sobre esse tema ganhe cada vez mais espaço no mundo acadêmico.

Como visto no desenvolvimento deste trabalho, a área de Informática é uma importante aliada tanto no ramo empresarial, para garantir vantagens competitivas nos meios de atuação de uma empresa, quanto no meio acadêmico auxiliando pesquisadores e servindo como uma importante ferramenta na comunidade cientifica para descoberta de conhecimento e análise dos dados.

Muitas ferramentas estão disponíveis para auxiliar o trabalho de pesquisadores de coletar, estudar e gerar conhecimento. Como exemplo dessas ferramentas pode-se citar mineração de dados e a geração de visualizações sobre as mais diversas informações.

A base de dados do INCA, estudada neste trabalho, possui informações valiosas para o estudo da doença no âmbito nacional. Mas essa base aliada a técnicas de visualização e disponibilizada em forma de uma aplicação fornece a ferramenta necessária para que até usuários com pouco conhecimento em informática possam explorar os dados sob uma nova ótica.

Porém, conforme apontado em resposta ao questionário enviado para especialistas da área, a análise destes dados utilizando ferramentas como o Excel torna a tarefa lenta e trabalhosa. A utilização da aplicação desenvolvida apresentou muitas vantagens para a interpretação dos dados da base. Respostas de um profissional da área da saúde ainda mostraram que a ferramenta é de fácil utilização, o que permite o uso também pelo público leigo.

Ainda há muito espaço para melhorias da ferramenta atual. Propostas de trabalhos futuros como melhoria da ferramenta existente com inclusão de novos atributos presentes na base, ampliação dos métodos estatísticos e implementação de tarefas de *data mining* na ferramenta podem ser exploradas.

Outra necessidade encontrada na execução deste trabalho e que abre outra alternativa de um trabalho futuro é a validação do conteúdo disponibilizado pelo INCA através do *dataset* utilizado. Conforme apontou o capítulo 4, dados dos pacientes de São Paulo continham registros inválidos para alguns atributos

que tiveram que ser corrigidos para apresentação na ferramenta. No decorrer da construção do trabalho, verificou-se que outros atributos como SEXO, RACACOR e LOCTUDET também continham inconsistências. Estes dados não foram alterados pois a quantidade de registros com problemas representava um volume muito pequeno de informação. Porém, a necessidade de uma melhor triagem e investigação da qualidade dos dados faz-se necessária para um aprofundamento maior no conhecimento da base de dados do INCA.

REFERÊNCIAS BIBLIOGRÁFICAS

AHLEMEYER-STUBBE, Andrea; COLEMAN, Shirley. **A Practical Guide to Data Mining for Business and Industry**. West Sussex, United Kingdom: John Wiley & Sons. 2014. p. 324.

AMERICAN CANCER SOCIETY. **The History of Cancer.** [S.l.], 2014. 14 p. SUDHAKAR, Akulapalli. **History of Cancer, Ancient and Modern Treatment Methods.** Nebraska: [s.n], 2009. 7 p. on-line.

ASAY, Matt. Data visualization: **Showing Isn't Always Telling**. 2016. Disponível em: < http://www.infoworld.com/article/3040708/analytics/data-visualization-showing-isnt-always-telling.html>. Acesso em: 08 de outubro de 2016.

CANCER.NET. Understanding Cancer Risk. 2016. Disponível em: < http://www.cancer.net/navigating-cancer-care/prevention-and-healthy-living/understanding-cancer-risk>. Acesso em: 05 de Setembro de 2016.

CHANG, Winston *et al.* **shiny: Web Application Framework for R** CRAN, 2017. Disponível em: < https://cran.r-project.org/web/packages/shiny/index.html >. Acesso em: 14 de Maio de 2017.

CHEN, Chun-houh; HÄRDLE, Wolfgang; UNWIN, Antony. **Handbook of Data Visualization.** 3º ed. German: Springer-Verlag BerlinHeidelberg, 2008. 899p.

CHIASSON, Trina; GREGORY, Dyanna. **Data + Design**: A simple introduction to preparing and visualizing information. [S.l.: s.n.]. Disponível em: < https://infoactive.co/data-design>. Acessado em: Acesso em: 08 de outubro de 2016.

DATAVIZCATALOGUE.COM. **The Data Visualisation Catalogue.** Disponível em: < http://datavizcatalogue.com/> Acesso em: 15 de outubro de 2016.

DEAN, Jared. **Big Data, Data Mining, and Machine Learning**. New Jersey, NY:John Wiley & Sons, 2014. p. 265.

ENCYCLOPEDIA BRITANNICA. **Cancer.** 2016. Disponível em: < http://global.britannica.com.secure.sci-hub.bz/science/cancer-disease>. Acesso em: 03 de Setembro de 2016. **********

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From Data Mining to Knowledge Discovery in Databases. **AI Magazine**, American Association for Artificial Intelligence. Menlo Park, California, EUA, 1996, v.17 n.3, p. 37-54.

FRIENDLY, M.; DENIS, D. J. (2001). **Milestones in the history of thematic cartography, statistical graphics, and data visualization**. Disponível em: <http://www.datavis.ca/milestones/>. Acesso em: 07 de setembro de 2016.

GOLDSCHMIDT, Ronaldo; PASSOS, Emmanuel. **Data Minining Um Guia Prático.** Rio de Janeiro: Elsevier Editora Ltda., 2005. 253p.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. **The Elements of Statistical Learning**: data mining, inference, and prediction. 2 ed. New York, NY: Springer-Verlag New York, 2009. p. 745.

IARC. **Cancer Today**. Lyon, France: International Agency for Research on Cancer. Cancer Today. Disponível em: <http://gco.iarc.fr/today>. Acesso em: 03 de Setembro de 2016.

INCA, **Atlas On-line de Mortalidade.** Disponível em: < https://mortalidade.inca.gov.br/MortalidadeWeb/pages/Modelo02/consultar.xhtml#panelResultado>. Acesso em: 21 de agosto de 2016.

KASPER, Dennis L. *et al*. **Harrison's Principles of Internal Medicine.** 19 ed. New York, NY: McGraw-Hill, v.2. 2015. p. 3000.

KATOUA, Hisham S. Exploiting the Data Mining Methodology for Cyber Security. **Egyptian Computer Science Journal.** Egito**,** Vol. 37 No. 6 September 2013. On-line.

MCCANDLESS, David. **The Beauty of Data Visualization.** 2010. Disponível em: < http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization?language=en>. Acesso em: 06 de Novembro de 2016.

MINISTÉRIO DA SAÚDE. **Estimativa/2016 Incidência de Câncer no Brasil.** Rio de Janeiro, 2015. 121p.

MISURA, Luciana. **Spring Break, o que é isso e como pode atrapalhar a sua viagem.** 2014. Disponível em: < http://luciana.misura.org/2014/02/20/spring-break-o-que-e-isso-e-como-pode-atrapalhar-a-sua-viagem/>. Acesso em: 06 de Novembro de 2016.

MUNZNER, Tamara. **Visualization Analysis & Design.** New York: CRC Press, 2014. 375p.

NATIONAL CANCER INSTITUTE. **Risk Factors for Cancer.** 2016. Disponível em: < https://www.cancer.gov/about-cancer/causes-prevention/risk>. Acesso em: 05 de Setembro de 2016.

ONCOGUIA. **Estimativas no Mundo.** 2015. Disponível em: < http://www.oncoguia.org.br/conteudo/estimativas-no-mundo/1706/1/>. Acesso em: 03 de Setembro de 2016.

ROSLING, Hans. **The Best Stats You've Ever Seen.** 2006. Disponível em: < http://www.ted.com/talks/hans_rosling_shows_the_best_stats_you_ve_ever_seen?language=en>. Acesso em: 06 de Novembro de 2016.

SAS. **Data Visualization Techniques**: From basics to bit data with SAS Visual Analytics. Disponível em: < http://www.sas.com/content/dam/SAS/en_us/doc/whitepaper1/data-visualization-techniques-106006.pdf>. Acesso em: 08 de outubro de 2016.

TAN, Pang – Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao DATAMINING Mineração de Dados.** Rio de Janeiro: Editora Ciência Moderna Ltda., 2009. 900p.

TEIXEIRA, Antonio C. B. *et al* **Câncer de Boca Noções Básicas para Prevenção e diagnóstico.** São Paulo: Editora Fundação Peirópolis, 1997. 86p.

WILLIAMS, Graham *et al.* rattle: **Graphical User Interface for Data Mining in R.** CRAN, 2016. Disponível em: < https://cran.r-project.org/web/packages/rattle/index.html>. Acesso em: 20 de Agosto de 2016.

World Health Organization, **Cancer.** Disponível em: <http://www.who.int/cancer/en/>. Acesso em: 21 de agosto de 2016.

YUK, Miko; DIAMOND, Stephanie. **Data Visualization For Dummies.** Hoboken, NJ: John Wiley & Sons. 2014. p. 256.

APÊNDICE A – Reposta Entrevistado 1

Área de Atuação do Respondente: Ensino e Pesquisa
Titulação do Respondente: Doutor

1) Você conhece a base de dados de câncer do INCA?

Sim

2) Você já utilizou a base de dados de câncer do INCA em alguma aula, estudo ou pesquisa científica?

Sim. Em aulas.

3) A ferramenta de visualização demonstrada pode auxiliar em uma melhor compreensão dos dados registrados na base de dados do INCA, por parte de gestores públicos, pesquisadores e usuários em geral?

Sim. Seria uma alternativa poderosa para a tomada de decisão de gestores. Pesquisadores podem, a partir da análise desses dados, visualizar padrões menos óbvios e delinear projetos e ações na área. Também, é uma ferramenta com manipulação bastante acessível pelo usuário.

4) Quanto ao tempo utilizado para analisar os dados na ferramenta de visualização, ela apresenta alguma vantagem/desvantagem em relação aos métodos tradicionais usados em seus estudos/pesquisas?

Na minha opinião, só apresenta vantagens. A possibilidade de visualizar dados robustos como estes de forma dinâmica pode economizar tempo do profissional de saúde, do gestor e de pesquisadores da área.

5) Quanto a velocidade de renderização dos gráficos e informações, qual a sua opinião sobre essa característica?

Apesar do primeiro “carregamento” de dados ser mais lento, não se compara ao tempo que seria necessário para a análise desses dados em formatos estáticos tradicionais.

6) Quanto a usabilidade da ferramenta de visualização, qual a sua opinião sobre essa característica?

A ferramenta é de fácil utilização, o que permitirá o uso também pelo público leigo. Análises complexas são disponibilizadas pela simples seleção dos parâmetros que se deseja conhecer.

7) Indique sugestões de avanços e melhorias na ferramenta de visualização.

Seria interessante se o menu lateral se movimentasse no sentido vertical sem que perdêssemos de vista o gráfico dinâmico (n 1)

Utilizar mais entradas da base de dados do INCA deixando o trabalho mais completo. Seria interessante a inclusão de atributos que constam na base do INCA e que possibilitariam a realização de mais análises. Informações importantes como as listadas abaixo:
DATAOBITO – data do óbito, dado para a determinação da taxa de mortalidade e taxa de cura (percentual de pessoas que morrem por conta da doença e que sobrevivem sem novo diagnóstico após cinco anos do término do tratamento).

DATAINITRT – data do início do tratamento, esse seria um dado importante para o conhecimento de taxa de pacientes que receberam e não receberam tratamento.

RZNTR – razão para não tratar, esse dado está diretamente relacionado ao fator acima, e seria interessante avaliar dos que não trataram, porque não trataram, e qual foi o desfecho, provavelmente óbito, quanto tempo sobreviveram sem o tratamento.

Mais algumas regressões seriam interessantes. Relação entre tabagismo e álcool com os diferentes escores de estadiamento do câncer. Relação entre tabagismo e álcool com o número de óbitos (dados que não aparecem no trabalho) e sobreviventes (dados que não aparecem no trabalho).

APÊNDICE B – Reposta Entrevistado 2

Área de Atuação do Respondente: Ensino e Pesquisa
Titulação do Respondente: Doutorado

1) Você conhece a base de dados de câncer do INCA?

Sim.

2) Você já utilizou a base de dados de câncer do INCA em alguma aula, estudo ou pesquisa científica?

Nunca, mas com esta ferramenta certamente irei utilizar.

3) A ferramenta de visualização demonstrada pode auxiliar em uma melhor compreensão dos dados registrados na base de dados do INCA, por parte de gestores públicos, pesquisadores e usuários em geral?

Sim. Acredito que a ferramenta resume de forma rápida e abrangente os dados, permitindo compreensão visual imediata a qualquer usuário, e em especial para os gestores, pode apontar rapidamente os dados que não têm sido preenchidos como rotina e são de grande interesse. Além disso, para pesquisadores, o delineamento de projetos fica muito facilitado com a ferramenta, que contempla análises iniciais para identificação de variáveis de interesse.

4) Quanto ao tempo utilizado para analisar os dados na ferramenta de visualização, ela apresenta alguma vantagem/desvantagem em relação aos métodos tradicionais usados em seus estudos/pesquisas?

Não identifiquei nenhuma desvantagem. A grande vantagem é a visualização de um grande banco de dados de forma dinâmica, permitindo ao usuário obter de forma extremamente rápida as informações contidas no banco, selecionando as variáveis de interesse de forma fácil. Os métodos tradicionais são muito dispendiosos em relação ao tempo, requerem auxílio de estatísticos ou usuários experientes, e programas robustos para suportar a quantidade de dados do INCA.

5) Quanto a velocidade de renderização dos gráficos e informações, qual a sua opinião sobre essa característica?

A velocidade é muito boa. Em trabalhos com bancos de dados muito menores em programas como o SPSS, por exemplo, a obtenção dos gráficos é mais lenta.

6) Quanto a usabilidade da ferramenta de visualização, qual a sua opinião sobre essa característica?

A ferramenta é extremamente 'user-friendly', permitindo seu uso por qualquer pessoa. A forma dinâmica com que a visualização pode ser manipulada de acordo com a seleção

das variáveis, permitindo análises estatísticas preliminares, torna sua usabilidade muito ampla.

7) Indique sugestões de avanços e melhorias na ferramenta de visualização.

Seria importante a inclusão de mais variáveis da base de dados do INCA, pois novas análises poderiam ser realizadas (regressão de Cox; Kaplan-Meier; tabelas de sobrevivência). Informações como:

DATAOBITO – data do óbito, dado para a determinação da taxa de mortalidade e taxa de cura (percentual de pessoas que morrem por conta da doença e que sobrevivem sem novo diagnóstico após cinco anos do término do tratamento). Com esta variável seria possível construir curvas de sobrevivência, inserindo variáveis diversas para avaliar a influência na sobrevida de determinado grupo.

DATAINITRT – data do início do tratamento, esse seria um dado importante para o conhecimento de taxa de pacientes que receberam e não receberam tratamento.

RZNTR – razão para não tratar, esse dado está diretamente relacionado ao fator acima, e seria interessante avaliar dos que não trataram, porque não trataram, e qual foi o desfecho, provavelmente óbito, quanto tempo sobreviveram sem o tratamento.

Na regressão, seria interessante informar, além do valor de r e r^2, o valor de p. Para o futuro, seria interessante a possibilidade de realizar regressão multivariada simples e ajustada, para que a estimação do efeito de 'tabagismo', 'álcool', 'idade', etc, possam ser ajustados.

APÊNDICE C – Script para migração dos arquivos do INCA

```
CREATE TABLE tb_inca (
ID SERIAL PRIMARY KEY,
TPCASO VARCHAR(20),
SEXO VARCHAR(20),
IDADE VARCHAR(20),
LOCALNAS VARCHAR(20),
RACACOR VARCHAR(20),
INSTRUC VARCHAR(20),
CLIATEN VARCHAR(20),
CLITRAT VARCHAR(20),
HISTFAMC VARCHAR(20),
ALCOOLIS VARCHAR(20),
TABAGISM VARCHAR(20),
ESTADRES VARCHAR(20),
PROCEDEN VARCHAR(20),
ANOPRIDI VARCHAR(20),
ORIENC VARCHAR(20),
EXDIAG VARCHAR(20),
ESTCONJ VARCHAR(20),
ANTRI VARCHAR(20),
DTPRICON VARCHAR(20),
DIAGANT VARCHAR(20),
BASMAIMP VARCHAR(20),
LOCTUDET VARCHAR(20),
LOCTUPRI VARCHAR(20),
TIPOHIST VARCHAR(20),
LATERALI VARCHAR(20),
LOCTUPRO VARCHAR(20),
MAISUMTU VARCHAR(20),
TNM VARCHAR(20),
ESTADIAM VARCHAR(20),
ESTADIAG VARCHAR(20),
RZNTR VARCHAR(20),
DTINITRT VARCHAR(20),
PRITRATH VARCHAR(20),
ESTDFIMT VARCHAR(20),
CNES VARCHAR(20),
UFUH VARCHAR(20),
MUUH VARCHAR(20),
OCUPACAO VARCHAR(20),
DTDIAGNO VARCHAR(20),
DTTRIAGE VARCHAR(20),
DATAPRICON VARCHAR(20),
DATAINITRT VARCHAR(20),
DATAOBITO VARCHAR(20),
OUTROESTA VARCHAR(20),
VALOR_TOT VARCHAR(20)
);

COPY tb_inca
(
TPCASO,
SEXO,
IDADE,
LOCALNAS,
RACACOR,
INSTRUC,
CLIATEN,
CLITRAT,
HISTFAMC,
ALCOOLIS,
```

```
TABAGISM,
ESTADRES,
PROCEDEN,
ANOPRIDI,
ORIENC,
EXDIAG,
ESTCONJ,
ANTRI,
DTPRICON,
DIAGANT,
BASMAIMP,
LOCTUDET,
LOCTUPRI,
TIPOHIST,
LATERALI,
LOCTUPRO,
MAISUMTU,
TNM,
ESTADIAM,
ESTADIAG,
RZNTR,
DTINITRT,
PRITRATH,
ESTDFIMT,
CNES,
UFUH,
MUUH,
OCUPACAO,
DTDIAGNO,
DTTRIAGE,
DATAPRICON,
DATAINITRT,
DATAOBITO,
OUTROESTA,
VALOR_TOT
)
FROM 'C:\\Temp\\inca\\db\\rhc85.csv'
DELIMITER ';'
CSV HEADER;
```

APÊNDICE D – *Queries* para verificação de inconsistência na base

```
SELECT dtpricon,sexo, count(1) from tb_inca
 where sexo not in ('1','2') group by dtpricon,sexo
2000  3     7
2001  3     3
2002  3     8
2003  3     16
2004  3     14
2005  3     14
2006  3     27
2007  3     21
2008  3     32
2009  3     49
2010  3     2
2011  0     2
TOTAL           195

SELECT dtpricon,racacor, count(1)
  from tb_inca where racacor not in ('1','2','3','4','5','9')
 group by dtpricon,racacor
2007  99    11
2008  99    124
2009  99    12
2010  99    10
2011  99    2
2013  99    2
2014  99    5
TOTAL           166

SELECT dtpricon,alcoolis, count(1) soma
  from tb_inca where alcoolis not in ('1','2','3','4','8','9')
 group by dtpricon,alcoolis
2000  0     1
2001  0     685
2002  0     1474
2003  0     1545
2004  0     2108
2005  0     2330
2006  0     4468
2007  0     6016
2008  0     7232
2009  0     6069
2010  0     467
2011  0     554
TOTAL           32.949

SELECT dtpricon,tabagism, count(1) soma
  from tb_inca where tabagism not in ('1','2','3','4','8','9') group by
dtpricon,tabagism
2001  0     685
2002  0     1475
2003  0     1545
2004  0     2108
2005  0     2330
2006  0     4468
2007  0     6016
2008  0     7232
2009  0     6069
2010  0     467
2011  0     554
TOTAL           32.949
```

```
SELECT dtpricon,loctudet, count(1) soma
  from tb_inca where loctudet not in
('C00','C01','C02','C03','C04','C05','C06','C07','C08','C09','C10','C11','
C12','C13','C14','C32','C15','C16','C17','C18','C19','C20','C21','C22','C2
3','C24','C25','C26','C30','C31','C33','C34','C37','C38','C39','C40','C41'
,'C42','C44','C47','C48','C49','C50','C51','C52','C53','C54','C55','C56','
C57','C58','C60','C61','C62','C63','C64','C65','C66','C67','C68','C69','C7
0','C71','C72','C73','C74','C75','C76','C77','C80','C78','C79','C81','C82'
,'C83','C84','C85','C86','C87','C88','C89','C90','C91')
  group by dtpricon,loctudet
1999  C     1
2000  C93   1
2000  C43   5
2000  C92   1
2002  C45   1
2002  C8    1
2003  C43   1
2005  C43   1
2005  C     2
2005  C46   3
2006  C43   2
2006  C     1
2007  C99   3
2007  C92   1
2007  C45   1
2007  C93   1
2007  C43   19
2008  C99   139
2008  C97   1
2008  C92   2
2008  C46   2
2008  C43   29
2009  C43   35
2009  C     1
2009  C92   3
2009  C99   7
2010  C43   26
2011  C96   1
2011  C43   9
2011  N62   1
2012  C43   14
2012  C45   1
2013  C43   10
2013  E05   1
2014  E05   1
2014  C96   1
2014  C45   1
2014  C43   8
TOTAL           338
```